(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2001－125929
（P2001－125929A）
(43)公開日　平成13年５月11日(2001.5.11)

(51)Int.Cl.[7]	識別記号	FI	テーマコード（参考）
G06F 17/30	170	G06F 17/30	170F
C12M 1/00		C12M 1/00	A
G01N 33/48		G01N 33/48	Z
G06F 3/00	651	G06F 3/00	651A
			651E

審査請求　未請求　請求項の数18　ＯＬ　外国語出願　（全 89 頁）　最終頁に続く

(21)出願番号　特願2000－227459(P2000－227459)
(22)出願日　平成12年７月27日(2000.7.27)
(31)優先権主張番号　０９／３６２６４９
(32)優先日　平成11年７月27日(1999.7.27)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　500014002
インサイト・ファーマスーティカルズ・インコーポレイテッド
ＩＮＣＹＴＥ　ＰＨＡＲＭＡＣＥＵＴＩＣＡＬＳ　ＩＮＣ.
アメリカ合衆国カリフォルニア州94304・パロアルト・ポータードライブ　3174
(72)発明者　アレックス・ジョージ・コレツァー
アメリカ合衆国　カリフォルニア州94560　ニューアーク，リンコナダ・コート，8260
(74)代理人　100096817
弁理士　五十嵐　孝雄　（外３名）
最終頁に続く

(54)【発明の名称】　生体分子配列データのためのグラフィカルビューア

(57)【要約】
【課題】　コンピュータベースの生体分子配列情報をグラフィック表示するための、方法、媒体、システムを開示する。
【解決手段】　生体分子配列情報は、一般に、本発明に従って種々様々な形式でグラフィック表示することができる。配列情報は、ヌクレオチド配列情報、アミノ酸配列情報、またはこれらの両方から構成される。配列情報は、配列に関して各々が異なる情報を提供するような複数の形式でグラフィック表示されて、コンピュータユーザインタフェースの１以上の画面に表示される。

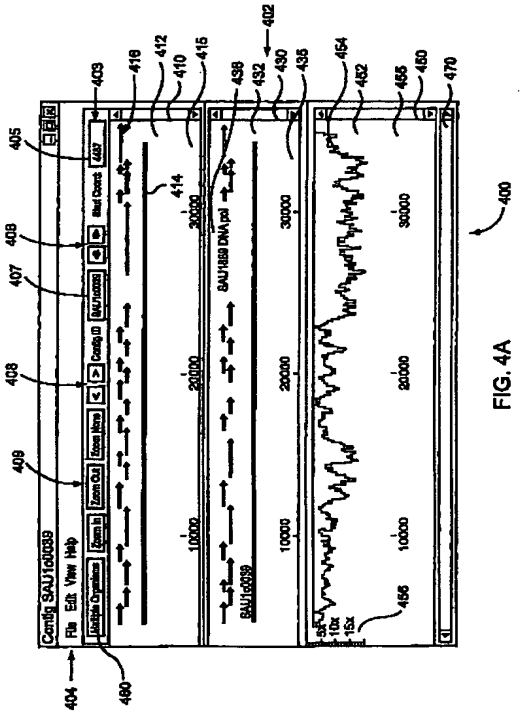


FIG. 4A

1

【特許請求の範囲】

【請求項１】　生体分子配列データを表わすためのコンピュータシステムにおいて実行される方法であって、
ユーザからの照会に応えて、データベースから生体分子配列データを検索し、
前記コンピュータシステムのユーザインタフェースにおいて、前記生体分子配列データのエレメントをグラフィック表示する、ことを備える方法。

【請求項２】　請求項１記載の方法であって、
前記グラフィック表示は複数のパネルを備え、前記複数のパネルの少なくとも１つは、前記複数のパネルのうちの少なくとも１つの他のパネルに表示された生体分子配列の一部位についてのカバレージ情報の深さをグラフィック表示する、方法。

【請求項３】　請求項２記載の方法であって、
前記複数のパネルは１つのフレーム内に設けられている、方法。

【請求項４】　請求項３記載の方法であって、
前記複数のパネルは、前記生体分子配列データの異なる側面を表わすグラフィック表示を提供する、方法。

【請求項５】　請求項４記載の方法であって、
前記生体分子配列データは遺伝子座位データである、方法。

【請求項６】　請求項５記載の方法であって、
前記複数のパネルは３つのパネルである、方法。

【請求項７】　請求項６記載の方法であって、
前記３つのパネルは、コンティグの少なくとも一部位とその関連座位とをグラフィック表示する第１のパネルと、前記第１のパネルに表示された前記コンティグの少なくとも一部位とその部位に関連した注釈付き座位とをグラフィック表示する第２のパネルと、前記第２のパネルに表示された前記コンティグの一部位のカバレージ情報の深さをグラフィック表示する第３のパネルと、を備える、方法。

【請求項８】　請求項７記載の方法であって、
前記第３のパネルは、前記第２のパネルに表示された前記コンティグの一部位のカバレージ情報の深さをグラフィック表示する、方法。

【請求項９】　請求項１記載の方法であって、
前記方法はＪａｖａプログラミング言語で実行される、方法。

【請求項１０】　生体分子配列データを表わすためのコンピュータシステムにおいて実行される方法であって、
ユーザからの照会に応えて、データベースから複数のホモロガス座位の生体分子配列データを検索し、
前記コンピュータシステムのユーザインタフェースにおいて、少なくとも幾つかの前記ホモロガス座位をグラフィック表示する、ことを備える方法。

【請求項１１】　請求項１０記載の方法であって、
前記グラフィック表示は１つのパネルを備える、方法。

2

【請求項１２】　コンピュータシステムであって、
生体分子配列データを含むデータベースと、
前記生体分子配列データに関する照会を受信し、前記照会の結果をグラフィック表示することが可能なユーザインタフェースと、を備えるコンピュータシステム。

【請求項１３】　請求項１２記載のシステムであって、
前記グラフィック表示は複数のパネルを備え、前記複数のパネルの少なくとも１つは、前記複数のパネルのうちの少なくとも１つの他のパネルに表示された生体分子配列の一部位についてのカバレージ情報の深さをグラフィック表示する、システム。

【請求項１４】　請求項１３記載のシステムであって、
前記複数のパネルは１つのフレーム内に設けられている、システム。

【請求項１５】　請求項１４記載のシステムであって、
前記複数のパネルは、前記生体分子配列データの異なる側面を表わすグラフィック表示を提供する、システム。

【請求項１６】　請求項１５記載のシステムであって、
前記生体分子配列データは遺伝子座位データである、システム。

【請求項１７】　請求項１６記載のシステムであって、
前記遺伝子座位データは、コンティグの少なくとも一部位とその関連座位とをグラフィック表示する第１のパネルと、前記第１のパネルに表示された前記コンティグの少なくとも一部位とその部位に関連した注釈付き座位とをグラフィック表示する第２のパネルと、前記第２のパネルに表示された前記コンティグの一部位のカバレージ情報の深さをグラフィック表示する第３のパネルと、を備える３つのパネルに表示される、システム。

【請求項１８】　生体分子配列データをグラフィック表示するようにアレンジされたプログラム命令を含むコンピュータ読み取り可能な媒体であって、
ユーザからの照会に応えて、コンピュータシステムデータベースから生体分子配列データを検索し、
前記コンピュータシステムのユーザインタフェースにおいて、前記生体分子配列データのエレメントをグラフィック表示する、ためのプログラム命令を含むコンピュータ読み取り可能な媒体。

【発明の詳細な説明】

【０００１】

【発明の背景】本発明は、概して生命情報学の分野に関し、特に、コンピュータベースの生体分子配列情報をグラフィック表示するための方法、媒体、およびシステムに関する。

【０００２】情報科学は、コンピュータおよび統計学的技術を情報管理に応用する学問である。生命情報学は、生体分子配列情報（例えば核酸やタンパク質等）のコンピュータデータベースを素早く検索する方法や、生体分子配列情報を分析して表示する方法、それにＤＮＡ配列データからタンパク質の配列、構造、および機能を予測

する方法の開発を含む。

【０００３】分子生物学は、実験室主体からコンピュータ主体へと急速に移行しつつある。今日の研究者にとって、配列と表現型の関係の探求には、計量分析やデータベース比較、計算アルゴリズムを高度に駆使できることが必須である。このため、遺伝子の配列や発現、そして分子構造の探求において、コンピュータ資源の使用を避けて通ることは不可能である。

【０００４】生命情報学の用途の１つは、生体のゲノムを研究することによって、その遺伝子の配列および配置を決定すること、ならびにその遺伝子が、同種ゲノム内の他の配列や遺伝子または他種の生体内の遺伝子に対して有する関連性を決定することである。こういった情報には、例えば薬効評価や薬物耐性評価へ応用するために、生物医学研究や薬学研究において大きな関心を寄せられている。ゲノム情報操作の実施および理解を容易にするために、高性能なコンピュータデータベースシステムも開発されている。米国カリフォルニア州パロアルトにあるインサイト製薬（Incyte Pharmaceuticals, Inc.）はそのようなデータベースをいくつか開発しており、その中には、ゲノム配列データが、電子的に記録され、また、公衆の配列データベースから得られる情報で注釈されたものもある。こういった公衆の配列データベースの例として、ＧｅｎＢａｎｋ（ＮＣＢＩ）やＳＷＩＳＳＰＲＯＴが挙げられる。結果として得られた情報は、リレーショナルデータベースに格納されて、同種のゲノム内または複数種のゲノム間で、配列と遺伝子の関連性を決定するのに利用される。

【０００５】遺伝データの処理や、インサイト製薬が開発したようなリレーショナルデータベースシステムが、遺伝情報の分析に多大な能力と柔軟性を提供している一方で、これらのシステムにさらなる改良を加えることにより、多数の用途において生物学的研究が加速されるであろう。

【０００６】対象となる分野の１つに、生体分子配列情報の表示が挙げられる。上述したように、ゲノム研究の重要な目的の１つは、生体遺伝子の配列および配置を決定すること、ならびにその遺伝子が、同種ゲノム内の他の配列や遺伝子、他種の生体内の遺伝子、または関連のタンパク質配列に対して有する関連性を決定することにある。与えられた１種または複数種の生体に対し、その遺伝子座位情報を明確且つ有効に表示することができれば、この課題に大きく貢献することができるに違いない。

【０００７】このように、与えられた１種または複数種の生体についての遺伝子座位情報および／または他の生体分子配列情報を、ユーザが明確かつ有効に表示させることのできる表示ツールの開発が望まれている。

【０００８】

【発明の概要】本発明は、コンピュータベースの生体分子配列情報をグラフィック表示するための、方法、媒体、システムを提供することにより、上記のニーズを満たすものである。生体分子配列情報は、一般に、本発明に従って種々様々な形式でグラフィック表示することができる。配列情報は、ヌクレオチド配列情報、アミノ酸配列情報、またはこれらの両方から構成される。配列情報は、配列に関して各々が異なる情報を提供するような複数の形式でグラフィック表示されて、コンピュータユーザインタフェースの１以上の画面に表示される。

【０００９】本発明のグラフィカルビューアは複数のパネルを備え、各パネルは対象とする生体分子配列データに関する情報をそれぞれ異なる方式で１画面または１ページ上に表示することが好ましい。第１のパネルは、例えば、生体分子配列全体またはその生体分子配列のうち対象とする一部位を、対象とする特定の部分配列の位置とともに、グラフィック表示することができる。第２のパネルは、例えば、第１のウィンドウに示された配列全体またはその選択部位を、より詳細にグラフィック表示することができ、これによって、ユーザは、対象とする特定部分配列に焦点を合わせることができる。この第２のパネルは、対象とした特定部分配列に関する注釈等の、付加的な情報を示すことができる。第３のパネルは、例えば、１以上の他パネルに示された生体分子配列データ等の信頼度や発生源を図示した情報を示すことができる。同じ画面または付加的な画面上の付加パネルは、例えば、１以上の他パネルに示された、選択された対象部分配列またはこれに関連する実際のヌクレオチド配列もしくはアミノ酸配列や、その生体分子配列データに関連する他の情報を、表示することができる。

【００１０】好ましい一実施形態において、本発明のグラフィカルビューアは、ある生体ゲノムの全体または選択部位を、その個々の座位とともにグラフィック表示することができる。このビューアによって、ユーザは、対象とする特定領域または特定座位に焦点を合わせ、それらを注釈等の付加情報とともにグラフィック表示することができるようになる。また、ビューアに示された配列領域に対し、その配列カバレージをグラフィック表示できるようにすることもできる。

【００１１】このビューアを使用すると、同種の生体ゲノムの他部位から得られる関連座位を表示することもでき、さらに、実際のヌクレオチド配列または詳細な注釈等の、その座位に関する情報を、関連のリレーショナルデータベースシステムから検索することもできる。また、本発明のグラフィカルビューアを使用すると、１つまたはそれ以上の生体ゲノムの複数部位を、対象座位ならびに対応するパラログおよびホモログ（別の生体ゲノムからの関連座位）に基づいて、グラフィック表示および比較することもできる。

【００１２】本発明の好ましい一実施形態に従ったグラフィカルビューアは、対象とするゲノムデータを、その

ゲノムデータに関して各々が異なる情報を表示するような複数のパネルを使用して、グラフィック表示することが好ましい。本発明の特に好ましい実施形態において、グラフィカルビューアは、１つの画面上にメインパネルを３つ有する。すなわち、考慮中のゲノム全体を示すレジェンドビューアと、特に関心のあるゲノム部位の領域にユーザが焦点を合わせる（ズームインする）ことを可能とするターゲットビューアと、考慮中のゲノム部位の長さに対してカバレージの深さを示したグラフィック情報を含む配列深さビューアと、の３つである。

【００１３】一態様において、本発明は、生体分子配列データを示すためのコンピュータシステムで実行される方法を提供する。この方法は、ユーザの照会に応えてデータベースから生体分子配列データを検索し、その生体分子配列データのエレメントをコンピュータシステムのユーザインタフェース内にグラフィック表示することを含む。この際、生体分子配列データを様々な側面から表示する複数のパネルが単一フレーム内に含まれているようにしてもよい。

【００１４】好ましい一実施形態において、生体分子配列データは、遺伝子座位データを含み、３つのパネル上にグラフィック表示される。すなわち、第１のパネルは、コンティグの少なくとも一部位とその関連座位とをグラフィック表示し、第２のパネルは、第１のパネル上に示されたコンティグの少なくとも一部位とその部位に関連した注釈付き座位とをグラフィック表示し、第３のパネルは、第２のパネル上に示された配列データを決定するために実施されたシークエンス操作の回数を示す情報をグラフィック表示する。第３のパネルは、第２のパネル上に示されたコンティグの一部位を構成するのに使用される配列か、または第２のパネル上に示されたコンティグの一部位に対するカバレージ情報の深さを、グラフィック表示することが可能である。

【００１５】別の一態様において、本発明は、生体分子配列データを示すためのコンピュータシステムで実行される別の方法を提供する。この方法は、ユーザの照会に応えて、データベースから複数のホモロガス座位の生体分子配列データを検索し、コンピュータシステムのユーザインタフェースにおいて、少なくとも幾つかのホモロガス座位をグラフィック表示することを含む。

【００１６】さらに別の一態様において、本発明はコンピュータシステムを提供する。このコンピュータシステムは、生体分子配列データを含むデータベースと、ユーザインタフェースと、を備える。このユーザインタフェースは、生体分子配列データに関連する照会を受信し、その照会結果をグラフィック表示することができる。

【００１７】さらにまた別の一態様において、本発明は、生体分子配列データをグラフィック表示するようにアレンジされたプログラム命令を含むコンピュータ読み取り可能な媒体を提供する。このコンピュータ読み取り可能な媒体は、ユーザの照会に応えて、データベースから生体分子配列データを検索し、コンピュータシステムのユーザインタフェースにおいて、その生体分子配列データのエレメントをグラフィック表示するためのプログラム命令を含む。

【００１８】以下の説明と、本発明の基本原理を実施例に従って示す明細書と添付の図面とから、本発明の上述したおよびその他の特徴および利点が一層明らかとなる。

【００１９】

【好ましい実施形態の詳細な説明】次に、本発明の好ましい実施形態について詳しく言及する。添付の図面には、好ましい実施形態が例示されている。以下では、本発明を好ましい実施形態と関連付けて説明するが、これは、本発明を１つまたはそれ以上の好ましい実施形態に限定することを意図するものではない。逆に、添付の特許請求の範囲によって定義される本発明の精神および趣旨の範囲内に含まれ得る、代替物、変更態様、および均等物をもカバーするものである。以下の説明では、本発明の完全な理解を促すために多くの項目を特定しているが、本発明は、これらの項目の一部または全てを特定しなくても実施することが可能である。そのほか、本発明を不必要に不明瞭化するのを避けるため、周知の処理工程の説明は省略した。

【００２０】導入：本発明は、コンピュータベースの生体分子配列情報をグラフィック表示するための、方法、媒体、システムを提供する。一般に生体分子配列情報は、本発明に従って種々様々な形式でグラフィック表示することが可能である。配列情報は、ヌクレオチド配列情報、アミノ酸配列情報、またはこれらの両方から構成される。配列情報は、配列に関して各々が異なる情報を提供するような複数の形式でグラフィック表示されて、コンピュータユーザインタフェースの１以上の画面に表示される。

【００２１】本発明のグラフィカルビューアは複数のパネルを備え、各パネルは対象とする生体分子配列データに関する情報をそれぞれ異なる方式で１画面または１ページ上に表示することが好ましい。第１のパネルは、例えば、生体分子配列全体またはその生体分子配列のうち対象とする一部位を、対象とする特定の部分配列の位置とともに、グラフィック表示することができる。第２のパネルは、例えば、第１のウィンドウに示された配列全体またはその選択部位を、より詳細にグラフィック表示することができ、これによって、ユーザは、対象とする特定部分配列に焦点を合わせることができる。この第２のパネルは、対象とした特定部分配列に関する注釈等の、付加的な情報を示すことができる。第３のパネルは、例えば、１以上の他パネルに示された生体分子配列データ等の信頼度や発生源を図示した情報を示すことができる。同じ画面または付加的な画面上の付加パネル

は、例えば、１以上の他パネルに示された、選択された対象部分配列またはこれに関連する実際のヌクレオチド配列もしくはアミノ酸配列や、その生体分子配列データに関連する他の情報を、表示することができる。

【００２２】好ましい一実施形態において、本発明のグラフィカルビューアは、ある生体ゲノムの全体または選択部位を、その個々の座位とともにグラフィック表示することができる。このビューアによって、ユーザは、対象とする特定領域または特定座位に焦点を合わせ、それらを注釈等の付加情報とともにグラフィック表示することができるようになる。また、ビューアに示された配列領域に対し、その配列カバレージをグラフィック表示できるようにすることもできる。

【００２３】このビューアを使用すると、同種の生体ゲノムの他部位から得られる関連座位を表示することもでき、さらに、実際のヌクレオチド配列または詳細な注釈等の、その座位に関する情報を、関連のリレーショナルデータベースシステムから検索することもできる。また、本発明のグラフィカルビューアを使用すると、１つまたはそれ以上の生体ゲノムの複数部位を、対象座位ならびに対応するパラログおよびホモログ（別の生体ゲノムからの関連座位）に基づいて、グラフィック表示および比較することもできる。

【００２４】本発明の好ましい一実施形態に従ったグラフィカルビューアは、対象とするゲノムデータを、そのゲノムデータに関して各々が異なる情報を表示するような複数のパネルを使用して、グラフィック表示することが好ましい。本発明の特に好ましい実施形態において、グラフィカルビューアは、１つの画面上にメインパネルを３つ有する。すなわち、考慮中のゲノム全体を常に示すレジェンドビューアと、特に関心のあるゲノム部位の領域にユーザが焦点を合わせる（ズームインする）ことを可能とするターゲットビューアと、考慮中のゲノム部位の長さに対してカバレージの深さを示したグラフィック情報を含む配列深さビューアと、の３つである。

【００２５】上述したように、本発明のグラフィカルビューアは、本発明の好ましい実施形態に関連して説明する遺伝子座位情報以外の生体分子配列情報の表示にも、もちろん使用することが可能である。本発明のグラフィカルビューアは、例えばペプチドまたはヌクレオチドの配列情報を表示するのに使用しても良く、例えばＢＬＡＳＴまたはＦＡＳＴＡを検索して配列を比較した結果として得られる、実際の配列を表示するのに使用することも可能である。

【００２６】グラフィカルビューアの環境：上述したように、本発明のグラフィカルビューアは、例えば、米国カリフォルニア州パロアルトにあるインサイト製薬（Incyte Pharmaceuticals, Inc.）によって開発され、また、米国特許第５，９７０，５００号、第５，９５３，７２７号、第５，９６６，７１２号、第６，０２３，６５９号などに記載されている生体分子配列のリレーショナルデータベースシステムとともに使用することが好ましい。本発明のグラフィカルビューアによって表示されるデータは、当業者が周知の技法およびコマンドを使用することにより、このようなデータベースから得られるものである。図１Ａおよび図１Ｂ、そして以下の関連する説明から、本発明のグラフィカルビューアの動作状況が分かる。

【００２７】図１Ａは、リレーショナルデータベースにおいて情報を格納および検索するのに適したネットワークシステム１３０を示す図であり、本発明のグラフィカルビューアをサポートするのに適したシステムである。ネットワーク１３０は、ネットワークサーバ１３６およびクライアント１３８ａ，１３８ｂ（より多数のクライアントの代表である）と、これらを接続するネットワークケーブル１３４とを備えている。このケーブル１３４は、ファイアウォールゲートウェイ１４０にも接続されており、ファイアウォールゲートウェイ１４０は、さらにインターネット１４２に接続されている。

【００２８】ネットワーク１３０は、ローカルエリアネットワーク（ＬＡＮ）や広域ネットワーク（ＷＡＮ）を含む周知の従来型ネットワークシステム（例えば、イーサネット（登録商標）やＩＢＭトークンリング等を使用したもの）のうちのいずれであっても良い。このネットワークは、周知のフォーマット（例えばＵＲＬ）のクライアントコールを任意のパラメータ情報とともに、ケーブルまたはワイヤ１３４を介して伝送するのに適した（１以上のパケットの）フォーマットにパッケージ化し、データベースサーバ１３６へ送達する機能を備えている。

【００２９】サーバ１３６は、ソフトウェアを実行させるのに必要なハードウェアを備えている。このソフトウェアは、（１）ユーザリクエストを処理するためのデータベースのデータにアクセスし、（２）クライアントマシン１３８ａ，１３８ｂに情報を与えるためのインタフェースを提供する。図１Ａの好ましい一実施形態において、サーバマシン上で実行されるソフトウェアは、サーバ－クライアント間でページデータを提供するためのワールドワイドウェブ（ＷＷＷ）プロトコルをサポートしている。本実施形態において、ＵＲＬおよびＨＴＴＰ機能を有するウェブサーバ１５６は、ＨＴＴＰプロトコルを介してクライアントとの通信を行う。

【００３０】クライアント／サーバ環境、データベースサーバ、リレーショナルデータベース、およびネットワークは、技術文献や、営業用文献、特許文献に詳細に記載されている。リレーショナルデータベースおよびクライアント／サーバ環境に関しては広く議論されており、特に、ＳＱＬサーバに関するデータベースサーバに関しては、例えば、「Nath, A., The Guide To SQL Server, 2nd ed., Addison-Wesley Publishing Co., 1995」を

参照すると良い。

【００３１】図示するように、サーバ１３６は、リレーショナルデータベース管理システム１５２と、ワールドワイドウェブアプリケーション１５４と、ワールドワイドウェブサーバ１５６とを実行させるオペレーティングシステム１５０（例えばＵＮＩＸ（登録商標））を備えている。サーバ１３６上のソフトウェアは種々の構成を想定でき、例えば、１つのマシン上に準備されていてもよいし、複数のマシンに分散して準備されていてもよい。

【００３２】ワールドワイドウェブアプリケーション１５４は、データベース言語ステートメント（例えば、スタンダードクエリーランゲージ（ＳＱＬ）ステートメント）の生成に必要な実行可能コードを含む。一般に、実行可能コードは埋め込みＳＱＬステートメントを含む。また、アプリケーション１５４は、コンフィギュレーションファイル１６０を含んでいる。このコンフィギュレーションファイル１６０は、サービスユーザリクエストにアクセスされていなければならない種々の外部および内部データベースと、サーバを構成する種々のソフトウェアエンティティと、に対するポインタおよびアドレスを含んでいる。また、コンフィギュレーションファイル１６０は、サーバ資源へのリクエストを適切なハードウェアに直接導く。これは、サーバが２以上の個別のコンピュータに分散している場合に必要となる。

【００３３】各クライアント１３８ａ，１３８ｂは、サーバ１３６へのユーザインタフェースを提供するためのワールドワイドウェブブラウザと、ＨＴＭＬページを生成するのに必要なコードとを備えている。各クライアント１３８ａ，１３８ｂは、ウェブブラウザを介して、例えば配列データベース１４４および／またはゲノムデータベース１４６からデータ検索するための検索リクエストを作成する。ユーザは、通常、ボタン、プルダウンメニュー、スクロールバー等の従来からグラフィカルユーザインタフェースとして採用されているユーザインタフェースエレメントを、指示したりクリックしたりする。クライアントのウェブブラウザでこのように作成されたリクエストは、ウェブアプリケーション１５４に伝送される。ウェブアプリケーション１５４は、リクエストのフォーマットを設定し、配列データベース１４４またはゲノムデータベース１４６から適切な情報を抽出するのに利用される照会を生成する。

【００３４】図中に示されている実施形態において、ウェブアプリケーションは、先ずデータベース言語（例えばＳｙｂａｓｅまたはＯｒａｃｌｅ　ＳＱＬ）で照会を作成することによって、ゲノムデータベース１４６のデータにアクセスする。このとき、このデータベース言語の照会が、リレーショナルデータベース管理システム１５２に引き渡され、そこで、データベース１４６から関連の情報を抽出できるように処理される。リクエストが配列データベース１４４にアクセスするものである場合、ウェブアプリケーション１５４は、データベース管理システム１５２のサービスを利用することなく直接そのデータベースにリクエストを通信する。

【００３５】ユーザリクエストにサービスする手順を、図１Ｂにさらに示す。本実施形態において、ワールドワイドウェブサーバ、および／または、サーバ１３６の実行可能なウェブアプリケーションコンポーネントは、クライアントマシンにハイパーテキストマークアップ言語文書（ＨＴＭＬページ）１６４を提供する。ＨＴＭＬ文書は、クライアントマシンにおいて、ユーザがデータベース１４６にアクセスするためのリクエストを作成するのに利用されるユーザーインタフェース１６６を提供する。このリクエストは、サーバ１３６のウェブアプリケーションコンポーネントによってＳＱＬ照会１６８に変換される。この照会は、サーバ１３６のデータベース管理システムコンポーネントによって使用されて、データベース１４６内の関連のデータにアクセスし、そのデータを適切なフォーマットでサーバ１３６に提供する。そして、サーバ１３６は、ウェブアプリケーション１５４を介して新しいＨＴＭＬ文書を生成し、データベース情報を、ユーザインタフェース１６６におけるビューアとしてクライアントに転送する。

【００３６】図１Ａで示される実施形態では、サーバ１３６とクライアント１３８ａ，１３８ｂの通信にワールドワイドウェブサーバおよびワールドワイドウェブブラウザを利用しているが、他の通信プロトコルを用いてもよい。例えばクライアントコールは、ＳＱＬ変換するウェブアプリケーション１５４に依らずに、ＳＱＬステートメントとして直接パッケージ化されていてもよい。クライアントは、クライアントブラウザを使用せず、直接データベースに照会することも可能である。

【００３７】ネットワーク１３０がワールドワイドウェブサーバおよびクライアントを利用する場合には、ＴＣＰ／ＩＰプロトコルをサポートしていなければならない。このようなローカルネットワークは、イントラネットと呼ばれることもある。このようなイントラネットの利点は、ワールドワイドウェブ上に属するパブリックドメインデータベース（例えばＧｅｎＢａｎｋワールドワイドウェブサイト）と容易に通信できることにある。本発明の特に好ましい一実施形態において、クライアント１３８ａ，１３８ｂは、ウェブブラウザおよびウェブサーバ１５６によって提供されたＨＴＭＬインタフェースを使用して、イントラネットデータベース上に属するデータに（例えばハイパーテキストリンクを介して）直接アクセスすることができる。

【００３８】ローカルデータベースの内容をプライベートに維持したい場合は、ファイアウォール１４０は配列データベース１４４およびゲノムデータベース１４６の内容を秘密に保たなくてはならない点に留意する必要が

ある。

【００３９】また、配列データベース１４４は、異なる種から得られるゲノム配列の単一ファイルを有したフラットファイルデータベースであるようにしてもよい。他には、配列データを、種別に、あるいは、ローカルデータベースに特有な配列（すなわち、ＧｅｎＢａｎｋ等の外部データベースで全く見つからなかった配列）であるか否かに基づいて、区分するようにしてもよい。

【００４０】ゲノムデータベース１４６内の情報は、リレーショナルフォーマットで格納されていることが好ましい。このようなリレーショナルデータベースは、関係代数によって定義される１組の操作をサポートする。リレーショナルデータベースは、一般に、同データベース内に含まれる各データに対し、列と行とで構成された複数のテーブルを含む。各テーブルは、テーブル内の行を一意的に識別する値を有した任意の列または列の組である基本キーを有する。リレーショナルデータベースのテーブルはまた、別のテーブルの基本キーの値に一致する値を有した列または列の組である外部キーを有する。リレーショナルデータベースは、一般に、データベース内の関係を管理する関係代数の基礎をなす１組の操作（選択する、射影する、作成する、結合させる、および分割する）に従う。上述したように、リレーショナルデータベースは、周知であり、また、多くの文献がある（例えば、前掲の「Nath, A., The Guide To SQL Server」を参照のこと）。

【００４１】リレーショナルデータベースは、種々の方法で提供される。例えばＯｒａｃｌｅ（商標）データベースでは、各テーブルにそれぞれ所有権が明記されたワークスペースがあるので、テーブルが物理的に分離されることはない。対照的に、Ｓｙｂａｓｅ（商標）データベースでは、異なるデータベース間でテーブルを物理的に分離させてもよい。

【００４２】複数のユーザに対応するネットワーク１３０の特定の構成では、１つのマシンにゲノムデータベースおよび配列データベースの双方を準備する。大きいな配列を検索する場合は、同サイズの第２のプロセッサを用意して、アプリケーションを２つのマシンに分割し、レスポンス時間を改善することが望ましい。

【００４３】デュアルプロセッササーバマシンとしては、Ｓｕｎ－Ｕｌｔｒａ－Ｓｐａｒｃ２（商標）（米国カリフォルニア州マウンテンビューにあるサンマイクロシステムズ）や、ＳＧＩ－Ｃｈａｌｌｅｎｇｅ Ｌ（商標）（米国カリフォルニア州マウンテンビューにあるシリコングラフィックス）、ＤＥＣ－２１００Ａ（商標）（米国マサチューセッツ州メイナードにあるデジタルエレクトロニクス）などのワークステーションが適している。マルチプロセッサシステム（４プロセッサ以上）としては、Ｓｕｎ－Ｕｌｔｒａ Ｓｐａｒｃ Ｅｎｔｅｒｐｒｉｓｅ ４０００（商標）や、ＳＧＩ－Ｃｈａｌｌｅｎｇｅ ＸＬ（商標）、ＤＥＣ－８４００（商標）などが適している。サーバマシンは、ネットワーク１３０のために構成されてＴＣＰ／ＩＰプロトコルをサポートしていることが好ましい。

【００４４】オペレーティングシステムは、利用されるワークステーションに依存しており、例えばＳｕｎ－Ｓｕｎ ＯＳ ５．５（Ｓｏｌａｒｉｓ ２５）、ＳＧＩ－ＩＲＩＸ ５３（またはそれ以上）やＤＥＣ－Ｄｉｇｉｔａｌ ＵＮＩＸ ３２Ｄ（またはそれ以上）を用いることができる。

【００４５】本発明に関連して使用されるデータベースは、４Ｘ４Ｇｂ＋ＦＷＳＣＳＩ－２、Ｆｉｂｅｒ Ｌｉｎｋ Ｒａｉｄ Ｕｎｉｔｓ ２ＯＧｂ＋、または４ＤＡＴ Ｔａｐｅ Ｄｒｉｖｅを介してダウンロードすることができる。また、ＣＤ－ＲＯＭドライブを用いることもできる。

【００４６】クライアントマシンとしては、例えば、Ｍａｃｉｎｔｏｓｈ（商標）（米国カリフォルニア州キューパティーノにあるアップルコンピュータ）や、ＰＣ、またはＵＮＩＸワークステーションを用いることができる。また、クライアントマシンは、Ｎｅｔｓｃａｐｅ（登録商標）やＩｎｔｅｒｎｅｔ Ｅｘｐｌｏｒｅｒ Ｗｅｂ Ｂｒｏｗｅｒを伴うＴＣＰ／ＩＰ対応であるべきである。

【００４７】ネットワークは、１０Ｂａｓｅ－Ｔや、１００Ｂａｓｅ－Ｔ、より高速の接続が可能なＴＣＰ／ＩＰ対応でもよく、外部のデータベースへのＨＴＭＬハイパーリンクのためにインターネットへのアクセスを提供する。

【００４８】図１Ｃは、本発明の好ましい一実施形態におけるグラフィカルビューア機能のアクセス可能性を示す図である。本発明のグラフィカルビューアは、ユーザが、生体分子リレーショナルデータベースのユーザインタフェースに表示されたユーザインタフェース画面の集合（例えばＨＴＭＬまたはＪａｖａ（登録商標）ページ）を介して利用できる１組の機能とともに提供されることが好ましい。通常、このインタフェースは、種々の照会ラインが続くメインのビューアページを有している。好ましい一実施形態において、メインビューアページ（およびその他のグラフィカルビューア）は、ネットワークシステム上で実行されるＪａｖａ（登録商標）ベースのアプレットである。当業者は、ここで説明される機能を与えられれば、本発明によるグラフィカルビューアを、Ｊａｖａ（登録商標）またはその他のプログラミング環境に実装することができる。ビューアページは、通常、グラフィカルビューアとともに使用される生体分子配列リレーショナルデータベースのユーザインタフェースの一部として提供される、別のページからアクセスされる。

【００４９】例えば、ユーザインタフェース画面（例え

ばＨＴＭＬページ）１７０は、複数の生体分子配列に関するテキスト情報を表示する。ページ１７０に表示された１以上の配列を、例えば、ＧＵＩとして設けられたポインタを使用して選択することにより、その選択された配列に関する付加的な情報を表示する別のページ１８０にアクセスする。このページ１８０に設けられたボタンを選択することにより、メインのグラフィカルビューアページ１９０にアクセスできる。グラフィカルビューアページ（例えばＪａｖａ（登録商標）ページ）１９０は、選択された配列に関する情報を図示する。このページには、ユーザによるグラフィカル表示の変更を可能とするための複数のボタン１９２を設けることが好ましい。ボタン１９２はさらに、付加的なグラフィカルビューアページ１９４，１９６にアクセスするための複数のボタンを備えていても良い。これらのページ１９４，１９６は、ページ１９０にグラフィック表示された配列情報に関する付加的な情報を、グラフィックまたは他の形式で表示する。

【００５０】遺伝子座位の提供：以下では、本発明を、遺伝子座位情報をグラフィック表示するための特に好ましい一実装形態と関連させて説明する。本発明は、米国特許第５，９７０，５００号に記述されるような、菌類データに適したデータベースに関連付けて説明されるが、本発明の適用範囲はこれに限定されない。例えば、本発明は、動物（例えばヒト、霊長類、げっ歯類、両生類、昆虫など）や植物の配列等の他の生体配列データに適したデータベースに関連して使用されるグラフィカルビューアをも範囲に含んでいる。

【００５１】上述したように、本発明のグラフィカルビューアは、ユーザが、生体分子リレーショナルデータベースのユーザインタフェースに表示されたユーザインタフェース画面の集合を介して利用できる１組の機能とともに提供されることが好ましい。メインビューアページは、通常、グラフィカルビューアとともに使用される生体分子配列リレーショナルデータベース（この場合は菌類ゲノムデータベース）のユーザインタフェースの一部として提供される、別のページからアクセスされる。図２は、菌類ゲノムデータベースから得られたこのような別のページの一例を示す。コンティグ結果ページ２００は、菌類の生体（この場合では黄色ブドウ球菌）のゲノム配列のうちの特定の「コンティグ」（重複配列の一集合体）、すなわちコンティグSAU1c0039 に局在化した遺伝子の座位（座位ＩＤによって識別される）のリストを表示している。

【００５２】ユーザは、コンティグ結果ページ２００に含まれる特定の座位ＩＤをクリックすることにより、図３に示されるような、座位情報ページにアクセスすることができる。例えばページ２００の座位ＩＤ：SAU100241 をクリックすると、その座位：SAU100241 に関する詳細を表示する座位情報ページ３００が返ってくる。このページは、選択されたときに本発明のグラフィカルビューアを起動するグラフィカルビューアボタン３０２を含んでいる。

【００５３】図４Ａは、座位情報ページ３００内のグラフィカルビューアボタン３０２を選択することによってアクセスされるメインのグラフィカルビューアページ４００を示している。この好ましい実施形態において、グラフィカルビューアは、あるコンティグの一部位およびその関連座位のグラフィック表示を提供するＪａｖａ（登録商標）ベースのアプレットである。本発明のグラフィカルビューアは、複数の個々のコンポーネントビューアを備えていることが好ましい。２以上のコンポーネントビューアが装備されている場合は、それらを１つのフレーム内に表示することにより、グラフィック表示されるデータをユーザに効率的に伝えることができる。好ましい実施形態では、３つのコンポーネントビューアが１つのフレーム内に表示される。

【００５４】ページ４００のグラフィカルビューア４０２は、１つの画面に３つのビューアコンポーネントパネルを有する。トップパネル４１０は、考慮中のゲノム全体を表示するレジェンドビューア４１２を備える。ミドルパネル４３０は、ユーザが特に関心のあるゲノム部位の領域に焦点を合わせる（ズームイン）ことを可能とするターゲットビューア４３２を備える。ボトムパネル４５０は、ターゲットビューア４２２に表示されているゲノム部位の長さに対して、カバレージの深さを示したグラフィック情報を表示する配列深さビューア４５２を備える。

【００５５】グラフィカルビューアページ４００は、付加情報へのアクセスおよびその表示のためのボタンおよびウィンドウを、ページ４００のトップ４０３に沿って幾つか有している。また、種々のコマンドおよび制御機能を一覧表示するプルダウンメニューにアクセスするためのメニューバー４０４が備えられている。各ビューアパネル４１０のボトムには、スケール４１５，４３５，４５５がそれぞれ示されている。これらの機能の使用方法に関しては、さらに後述する。

【００５６】レジェンドビューア４１２は、ユーザが前の画面で１つのコンティグを選択した際にビューアによってロードされたそのコンティグの全部位を常に示している。好ましい一実施形態において、ビューアは、そのコンティグ配列に関して所定のデフォルト数分の塩基対をロードする。コンティグがデフォルトより短い場合は、コンティグ全体が表示されてデフォルトが調整される。例えば本実施形態において、ビューアは、コンティグ結果画面２００上のリストに示される第１の座位、すなわちg2462967（ヒットＩＤによって識別される）からスタートし、合計３０，０００の塩基対をロードする。示される塩基対の番号およびコンティグ上のその位置は、レジェンドビューアパネル４１０のボトムに示され

るスケール４１５をもとに決定される。デフォルト値は、もちろん任意の望ましい数に変更することが可能である。

【００５７】レジェンドビューア４１２は、スケール４１５からわかるように、コンティグSAU1c0039 を、座標（塩基対番号）４４６７を起点とし座標３４，４６７に至るライン４１４としてグラフィック表示している。ビューアに示されたコンティグは、コンティグＩＤウィンドウ４０７内で識別される。また、レジェンドビューア４１２によって示されるコンティグ部位の開始座標（すなわち、選択された座位g2462967の開始座標：４４６７）は、開始座標ウィンドウ４０５に示されている。これらのウィンドウ４０５，４０７は、以下で説明するように、ビューアによって表示される情報を制御するための情報を入力するのに使用してもよい。ユーザは、方向ボタン４０６をクリックすることにより、レジェンドビューア４１２および他のコンポーネントビューアで、コンティグの上流部位または下流部位を見ることができる。

【００５８】レジェンドビューア４１２は、コンティグに加え、コンティグの部位上にある種々の座位をも表示する。これらの座位を表示する方法から、本発明のグラフィカルビューアが、効率的な情報表示において力を発揮することがわかる。

【００５９】座位は、矢印４１６で表わされる。各座位は、そのコンティグ上の位置およびそれが読まれる方向に従って、コンティグライン４１４に沿って配置されている。矢印は、その座位が読まれる方向を表わしている。前進方向（＋）に読まれる座位はコンティグライン４１４の上方に示され、後退方向（－）に読まれる座位はコンティグライン４１４の下方に示される。また、グラフィック表示された座位に関する情報を伝えるために、他のグラフィック機能を使用してもよい。例えば、確定された信頼性の閾値を上回る配列の元となる座位は、破線の矢印で示すことが可能である。

【００６０】この好ましい実施形態では、タンパク質機能に基づいて座位を色分けする。タンパク質を種々の機能カテゴリにグループ分けし、各カテゴリに色を割り当てる。例えば、この好ましい実施形態では、座位に対応するタンパク質は以下のカテゴリ／色に従ってグループ分けされる。すなわち、運動性／淡青色、病原性／赤色、輸送性／黄緑色、制御性／マゼンタ、高分子の新陳代謝／黄色、低分子の新陳代謝／濃青色、構造／濃緑色、未分類／黒である。もちろん、他のカテゴリおよび色を使用してもよい。座位の機能を表示するこれら矢印や色分けは、レジェンドビューアと以下で説明するターゲットビューアとの両方で使用される。

【００６１】ターゲットビューア４３２は、初めはレジェンドビューア４１２と同じ範囲を表示するが、ズームボタン４０９をクリックすることにより、範囲を変化させることが可能である。「Ｚｏｏｍ　Ｉｎ」ボタンを使用すると、レジェンドビューア４１２に示されたコンティグの部位を詳細表示することができる。そしてターゲットビューア４３２には、ズーム倍率に応じて調整されたスケール４３５を伴う詳細表示が示される。「Ｚｏｏｍ　Ｏｕｔ」ボタンを使用すると、最大でレジェンドビューアのために選択されたデフォルトの塩基対数（最小倍率）まで、コンティグを広範囲に表示することができる。「Ｚｏｏｍ　Ｎｏｎｅ」ボタンを使用すると、自動的に最小倍率に戻る。

【００６２】本発明のグラフィカルビューアによって提供されるレジェンドビューア４１２に示されたコンティグ４１４の対象部位に焦点を合わせるための別の方法としては、コンティグ４１４の対象部位の周囲にカラー（例えば赤色）ボックス等のアウトラインを提供し、それをターゲットビューア４３２に表示する方法が挙げられる。この好ましい実施形態において、赤色ボックスがレジェンドビューアパネル全体を囲んだ場合には、ターゲットビューアは３０，０００の塩基対を全て表示する。これが、図４Ａに示された状態である。上述したように、赤色ボックスはズームボタン４０９を用いて調整される。

【００６３】ユーザは、赤色ボックス（ラバーバンディングとして知られる）を直接調整することによっても、コンティグ上の一領域を拡大・縮小することが可能である。赤色ボックスの範囲の変更は、任意のビューアパネルで１つの場所をクリックし、カーソルをマウスとともに別の場所にドラッグすることによって行われる。このとき、赤色ボックスがこれら２点に挟まれる領域を取り囲み、ターゲットビューアおよび配列深さビューアでは、この領域のみが見られるようになる。図４Ｂは、ユーザが、レジェンドビューア４１２に示されるコンティグ４１４のうち、座標が約１４，２００から１８，２００までの部位４３４を拡大したときのビューア４０２の更新ページを示したものである。ターゲットビューア４３２のボトムにあるスケール４３５は、拡大されたターゲット表示の新しい範囲に応じて調整されている。

【００６４】ターゲットビューアの他の機能は、座位が注釈されていることである。図４Ａおよび４Ｂで見られるように、注釈情報を載せるのに十分な長さの座位の場合は、座位の矢印に注釈４３６が付される。対象座位が注釈を載せるのに短すぎる場合には、ユーザは、注釈のグラフィック表示に十分な長さまで、その座位を拡大すれば良い。

【００６５】ターゲットビューア４３２にグラフィック表示された座位をクリックして選択することにより、個々の座位に対して更なる分析をすることができる。選択された座位は、例えばカラー（例えば赤色）ボックスで囲む等の何らかの方法によって強調される。そして、座位の表示をダブルクリックすることによって、その座位

の詳細を見ることができる。つまり、ダブルクリックによって、図５Ａに示されるような、選択オブジェクト詳細ウィンドウが開く。選択オブジェクト詳細ウィンドウ５００は、座位に関する情報を含んでいる。この情報には、座位ＩＤ、遺伝子（機能別）カテゴリ、塩基対の範囲、例えばＧｅｎＰｅｐｔデータベース等の他の配列データベースに対する配列のホモロガスマッチ（返されるホモロガスマッチ数は、例えば上位５つ等の予め設定された数に制限されることが好ましい）、検索者にとって有用でかつグラフィカルビューアとともに使用されるデータベースシステムの他の機能に関連するその他の情報などが含まれる。ウィンドウ５００で提供される情報フィールドの多くは、他のＨＴＭＬページまたは他の画面へのハイパーリンクであってもよい。

【００６６】選択オブジェクト詳細ウィンドウ５００は、アライメントボタン５０２を備える。このボタンをクリックすると、座位の配列およびそのホモロガスマッチをグラフィック表示するアライメントビューアにアクセスすることができる。図５Ｂは、本発明の好ましい一実施形態におけるアライメントビューア５１０の一例を示したものである。アライメントビューア５１０は、パネルを３つ有する。上２つのパネル５１２，５１４は、図５Ａに示される座位（SAU101156 ）をグラフィック表示する。３番目のパネル５１６は、図５Ａで記された５つのホモログをグラフィック表示する。アライメントビューアページは、グラフィック表示の制御に使用されるいくつかのボタン５１８を含んでいる。このページは、特に、下２つのパネル５１４，５１６に示される座位を配列レベルで拡大・縮小するのに使用されるズームボタン５２０を備える（一方、一番上のパネル５１２は、座位全体を表示したままである）。図５Ｃは、このようなズーム機能を示したものであり、上のパネル５１２では、下２つのパネル５１４，５１６に、ホモログとともに配列レベルで示された座位の一部位が、カラーボックス５２２で囲まれている。この実施形態ではアミノ酸配列が示されているが、他の実施形態では、対応するヌクレオチド配列を示すことも可能である。

【００６７】グラフィカルビューアページ４００が有する他の機能のうち、同ページのボトムに設けられたスクロールバー４７０は、ターゲットビューア４３２の表示範囲を、レジェンドビューア４１２に示されるコンティグ配列の一部位に焦点合わせした場合に有用となる。ユーザは、スクロールバー４７０をコンティグ４１４の一部位４３４に沿って移動させることにより、ターゲットビューア４３２の表示範囲に、同コンティグ４１４の上流または下流の部位を表示させることができる。

【００６８】本発明のこの実施形態において、グラフィカルビューア４０２の３番目のパネル４５０は、配列深さビューア４５２である。配列深さビューア４５２は、カバレージの深さ、すなわちコンティグの所定部位がシークエンシングされた回数を、同コンティグの長さに対して示したグラフを表示する。配列深さビューア４５２は、ターゲットビューア４３２に表示されたコンティグまたはコンティグの一部位を対象としたグラフを表示する。このため、ターゲットビューア４３２とレジェンドビューア４１２が同じ範囲を示す図４Ａの場合には、配列深さビューア４５２が表示するグラフは、配列深さビューアパネル４５０のボトムに位置するスケール４５５が示すように、カバレージの深さを、コンティグ４１４のうち座標４４６７～３４，４６７に相当する３０，０００塩基対に対して示したものである。一方、図４Ｂの場合には、配列深さビューア４５２が表示するグラフは、調整後のスケール４５５が示すように、カバレージの深さを、コンティグのうちターゲットビューアでクローズアップされている座標約１４，２００～１８，２００に相当する部位４３４の約４０００塩基対に対して示したものである。配列深さビューアは、さらに、グラフに表示されたシークエンシングの通過回数を示す第２のスケール４５６をｙ軸上に備えている。

【００６９】このカバレージ情報の深さを示す方法から、本発明のグラフィカルビューアが効率的に情報表示できることがわかる。グラフィカルビューアのユーザは、一見するだけで、ビューア中の他のパネルに示された配列情報の信頼性に関する有用な情報を、素早く知ることができる。本発明のこの好ましい実施形態における配列深さビューア４５２は、カバレージを配列分布グラフ４５４として表示する。カバレージ情報の深さをこのように表示すれば、この情報を、ｙ軸スケール４５６を参照することにより、視覚的に明確な印象を与えかつ容易に定量が可能なグラフ形式で、特に効率的に提供できる。種々の領域のカバレージデータも、この形式によって容易に比較することができる。

【００７０】本発明の別の実施形態では、配列深さビューアは、カバレージ情報の深さを別の方法でグラフィック表示してもよい。例えば、コンティグが構成される実際の配列を表示してもよい。この方法で配列のカバレージ情報を表示すれば、データ収集プロセスに関心のあるユーザにとって有用な情報、例えばコンティグの形成に使用される情報を提供することができる。

【００７１】上述したように、グラフィカルビューアページ４００は、同ページ４００のトップ４０３に沿って、付加的な情報へのアクセスおよびその表示を目的としたいくつかのボタンを備える。これらのうち、開始座標４０５ウィンドウやコンティグＩＤ ４０７ウィンドウ等の幾つかに関しては、すでに説明済みである。図６および７は、本発明のグラフィカルビューアのこの実施形態における付加的な機能を示したものである。

【００７２】開始座標ウィンドウ６０５は、レジェンドビューア６１２に表示されたコンティグ配列の開始座標を表示するのに加え、ユーザから異なる開始座標の入力

を受け付けることも可能である。異なる開始座標を入力すると、コンティグ配列の異なる部位がレジェンドビューアに表示される。例えば、図６に示されるグラフィカルビューアページ６００では、開始座標 ウィンドウ６０５に０が入力されている点を除き、ページ４００と同じように設定されている。このため、レジェンドビューア６１２に示されるコンティグ配列６０２および関連の座位６０４が、４４６７塩基対分だけ上流に推移して、コンティグSAU1c0039 の先頭にきている。図４に示されるコンティグ４１４の最下流４４６７に位置する塩基対は、ページ６００のビューアではもはや見ることができない。また、ターゲットビューア６３２および配列深さビューア６５２には、対応する表示がなされている。

【００７３】また、コンティグＩＤウィンドウ４０７は、ビューア４０２に表示されているコンティグを識別するのに加え、ユーザから異なるコンティグＩＤの入力を受け付けることも可能である。異なるコンティグＩＤの入力によって、新しいコンティグＩＤに関連付けられたコンティグ配列のうちデフォルト数分の塩基対（座標０からスタートすることが好ましい）が、ビューアに関連付けられたデータベースからロードされて表示される。図７は、例えばコンティグＩＤウィンドウ７０７にコンティグＩＤ：SAU1c0016 を入力した際のグラフィカルビューアページ７００を示している。このとき、レジェンドビューア７１２には、コンティグSAU1c0016 のコンティグ配列７０２およびその関連座位７０４が表示されている。また、ターゲットビューア７３２および配列深さビューア７５２には、対応する表示がなされている。

【００７４】また、上述したように、グラフィカルビューア４００は、種々のコマンドおよび制御機能を一覧表示するプルダウンメニューにアクセスするためのメニューバー４０４を備える。「Ｆｉｌｅ」プルダウンメニューには、保存や印刷等の、アプリケーションソフトウェアパッケージに見られる標準的なコマンドが一覧表示されている。「Ｅｄｉｔ」プルダウンメニューは、コンティグ配列長さのデフォルト表示数や、ビューア内で種々の特徴を表示するのに使用される色などの、グラフィカルビューアのパラメータを編集するためのカテゴリリストを提供する。

【００７５】特に重要なのは、「Ｖｉｅｗ」プルダウンメニューである。この「Ｖｉｅｗ」プルダウンメニューは、ユーザが種々のビューア表示に含ませる機能を選択できるようにするとともに、座標上の配列表示８０４オプションをも備えている。図８Ａは、「Ｖｉｅｗ」プルダウンメニューが選択された状態のグラフィカルビューアページ８００を示したものである。メニュー８０２において、座標上の配列表示８０４を選択すると、グラフィカルビューア４０２に示されたコンティグを構成するのに使用される配列が、その各配列のカバレージを示す座標とともに一覧表示されたページ８１０にアクセスすることができる。このリストから配列を１つ（例えば２番目の806503054F1(5201,5690)８１２）選択し、配列データベースボタン８１４をクリックすると、グラフィックビューアに関連付けられたデータベースシステムであってコンティグを構成するのに使用される、未加工の配列データベースにアクセスする。このとき、図８Ｃに示されるような、配列検索結果ページ８２０が返される。配列検索結果ページ８２０は、図８Ｂで選択された配列８１２の実際のヌクレオチド配列８２２を示す。

【００７６】図９は、本発明の好ましい一実施形態におけるグラフィカルビューアシステムが、遺伝子座位情報をユーザに返す際のプロセスを一般化して示すフローチャートである。このフローチャートでは、本発明の一実施形態における配列データをグラフィック表示するためのオプションをフローチャートの形式で示すために、本発明の好ましい一実施形態における主な機能のみが選択して示されている。よってこのフローチャートは、本発明を包括的に示すことを意図したものではない。

【００７７】プロセス９００がステップ９０１でスタートすると、ステップ９０２において、選択された座位に対応するデータおよびその関連のコンティグがグラフィカルビューアにロードされる。上述したように、座位は、グラフィカルビューアとともに使用される生体分子配列リレーショナルデータベース（この場合は細菌ゲノムデータベース）のユーザインタフェースの一部として提供される、ＨＴＭＬページ内のリストから選択することができる。ステップ９０４では、コンティグ上の選択された座位がグラフィック表示される。このグラフィック表示は、選択座位に関連する生体分子配列データを異なる種々の側面から表示するための複数のコンポーネントを有することが好ましい。図４Ａおよび４Ｂに示された特に好ましい実施形態では、レジェンドビューアと、ターゲットビューアと、配列深さビューアとの３つのコンポーネントを有するビューアによってグラフィック表示される。

【００７８】さらなる入力またはズーム調整がなされなかった場合、上記プロセスは、ステップ９０４におけるグラフィック表示に続いてステップ９４０で終了する。しかしながら、ユーザがグラフィカルビューアを使用して選択座位またはその他座位に関する付加情報を抽出および表示したい場合には、ビューアは、その目的に合わせて付加的な機能を提供することができる。

【００７９】グラフィカルビューアによって表示されたデータのグラフィック表示は、種々の方法で変更することが可能である。また、ビューア内の種々のオブジェクト（すなわち座位）を選択することによって、付加的な情報にアクセスすることも可能である。例えば、ユーザは、グラフィカルビューアページ内に設けられた一領域、例えば図４Ａのウィンドウ４０７に、新しいコンテ

ィグＩＤを入力しても良い。すると、判断ステップ９０６が肯定応答し、ステップ９１４において、ビューアに新しいコンティグおよびその座位がグラフィック表示される。また、ユーザは、例えば図４Ａの開始座標ウィンドウ４０５に、新しい開始座標を入力しても良い。すると、判断ステップ９０８が肯定応答し、ステップ９１６において、グラフィック表示が調整されて新しい座標範囲のコンティグが表示される。上述したように、ユーザはさらに、グラフィック表示されたコンティグの特定部位に焦点を合わせることも可能である。このとき、判断ステップ９１０が肯定応答し、ステップ９１８において、本実施形態ではターゲットビューアのグラフィック表示が調整されて、そのコンティグが拡大表示される。上述したどの判断ステップにおいても、否定応答された場合にはグラフィカルビューアの表示は変化しないまま維持される。

【００８０】上述したアクションの後かまたはそれに替わり、ユーザは、さらなる情報を得る目的でオブジェクトを選択することも可能である。好ましい一実施形態において、グラフィカルビューアのターゲットビューアコンポーネントに表示された座位は、その表示をクリックすることによって選択することが可能である。すると、判断ステップ９２０が肯定応答し、ターゲットビューアにおけるその座位の表示がカラーボックスで強調される。選択された座位に関して詳細な情報を得たい場合は、ユーザは、その座位の表示をダブルクリックする。すると、判断ステップ９２１が肯定応答し、ステップ９２２において、選択座位に関する詳細な情報を示したＪａｖａ（登録商標）ページが表示される。

【００８１】本発明の好ましい一実施形態における他の特徴は、上述したように、グラフィカルアライメントビューアを備えることである。ユーザは、対象座位のアミノ酸配列の配置を幾つかのホモロガス配列に対して示すグラフィカルビューアの表示を選択することが可能である。すると、判断ステップ９２４が肯定応答し、ステップ９２６において、そのアライメントがグラフィカルビューアにグラフィック表示される。

【００８２】ユーザはまた、対象座位のホモロガス座位およびパラロガス座位のグラフィック表示を見るために、マルチプル生体ビューアの表示を選択することも可能である。例えば、判断ステップ９２０が肯定応答し、さらに判断ステップ９２８が肯定応答すると、ステップ９３０において、マルチプル生体ビューアにアクセスすることが可能となる。本発明の好ましい一実施形態におけるマルチプル生体ビューアの操作に関しては、図１０Ａ～１０Ｅおよび図１１とを参照しつつ、以下でさらに詳述する。

【００８３】もちろん、選択されたオブジェクトの詳細を表示する選択判断と、マルチプル生体ビューアを表示する選択判断とは互いに独立しており、図９以外の方法でも容易に可能である。なお、このシステムでは、ユーザがグラフィカルビューアモードをいつでも終了できるようになっているが、この選択肢は図９には示されていない。

【００８４】本発明のグラフィカルビューアからさらなる情報にアクセスするために使用可能なさらなる選択肢の１つに、対象座位とそのコンティグとに関連付けられた選択配列の実際のヌクレオチドまたはアミノ酸配列を表示する選択肢がある。好ましい一実施形態において、ユーザは、図４Ａに示すようなグラフィカルビューアページのボタンをクリックすることによって、この選択肢を選択することが可能である。このとき、判断ステップ９１２が肯定応答し、ステップ９３２において配列のリスト（配列の識別子、および、ビューアに表示されたコンティグを構成する配列の座標）が表示される。そして、ユーザがそのリストから配列を１つ選択すると、ステップ９３２が肯定応答し、選択された配列の実際の（この場合は）ヌクレオチド配列が表示される。そして、ステップ９４０でプロセスが終了する。

【００８５】本発明のグラフィカルビューアに表示される他のデータと同様に、本発明のこの態様で使用されるデータも、関連の生体分子配列データベースおよびシステムから得られる。このようなシステムの編成および操作は多様であり、本明細書で前掲したインサイト製薬の特許にその例が記述されている。当業者ならば、本明細書に記載された機能および表示の説明をもとにして、このようなシステムに本発明によるグラフィカルビューアを実装することが可能である。

【００８６】マルチプル生体ビューア：上述したように、本発明のグラフィカルビューアを用いれば、１つまたはそれ以上の生体のゲノムに含まれる複数の部位を、対象座位ならびにそれに対応するパラログ（同じ生体のゲノムの他部位から得られる関連の座位）およびホモログ（別の生体のゲノムから得られる関連の座位）に基づいて、グラフィック表示および比較することが可能である。このようなマルチプル生体ビューアの好ましい一実施形態を、図１０Ａ～１０Ｄを参照しつつ説明する。

【００８７】図１０Ａは、図４Ａおよび４Ｂに示されるような、メインのグラフィカルビューアページ１０００を示したものである。図１０Ａでは、ボックス（ラバーバンド）１００２が、グラフィカルビューア１０１０のレジェンドビューア１００８コンポーネントにおいて表示されているコンティグ１００６部位の一領域１００４を取り囲むように配置されている。コンティグ１００６のこの領域１００４は、グラフィカルビューア１０１０のターゲットビューア１０２０コンポーネントに表示されており、そのカバレージは、配列深さビューア１０３０コンポーネントに表示されている。ターゲットビューア１０２０において、座位SAU100242 を囲むボックス１０２２は、その座位が選択されたことを示している。前

述したように、メインビューアページ１０００は、マルチプル生体ボタン１００１を備えている。

【００８８】ターゲットビューアで座位が選択された際に、マルチプル生体ボタン１００１をクリックすると、ビューアと関連付けられたデータベースが検索され、選択座位のホモログおよびパラログを含む全ライブラリが一覧表示される。図１０Ｂは、図１０Ａのページで選択された座位SAU100242 をもとに検索されたライブラリリストを示すウィンドウ１０４０である。ホモログおよびパラログの個々のリストにアクセスするためには、ユーザは、このウィンドウ１０４０に表示されたリストから１つまたはそれ以上のライブラリを選択すれば良い。そして、さらにマルチプル生体ボタン１０４２をクリックすると、個々のホモログおよび／またはパラログが検索されて表示される。図１０Ｃは、図１０Ｂの画面１０４０において選択されたライブラリから得られる座位SAU100242 のホモログおよびパラログのリストを示すウィンドウ１０４５の一例を示したものである。

【００８９】次に、ユーザは、初めに選択した座位（例えばSAU100242 ）、ならびに図１０Ｃのリストに示される選択されたホモロガス座位およびパラロガス座位を、グラフィック表示するよう選択することが可能である。ウィンドウ１０４５のマルチプル生体ボタン１０４６をクリックすると、対象座位、ならびにそのホモログおよびパラログが、本発明の好ましい一実施形態におけるマルチプル生体ビューアにロードされて、対象座位、ならびに選択されたホモログおよびパラログが表示される。図１０Ｄは、このようなマルチプル生体ビューアページ１０５０の一例を示したものである。

【００９０】マルチプル生体ビューアページ１０５０は、コンティグ（SAU1c0039 ）上の選択された対象座位（SAU100242 ）、ならびにコンティグ上の選択されたホモロガス座位およびパラロガス座位をグラフィック表示する１つのパネルで構成されたマルチプル生体ビューア１０５２を提供する。図１０Ｄに示すビューア１０５２は、１つのページに５つのコンティグ：SAU1c0039 １０６１、PRT1c0129 １０６２、SAU2c0391 １０６３、SEP1c0220 １０６４、SHA1c0122 １０６５をグラフィック表示している。コンティグ１０６１は、その座位とともに示されており、より明確に識別できるように太字イタリック体で表示された選択座位SAU100242 １０７１を含んでいる。図１０Ｄに示されている一実施形態では、座位を、座位ＩＤよりもむしろヒットに関する記述によって注釈している。他の各コンティグもまた、横に沿って示された座位、ならびに太字イタリック体で示されたSAU100242 にホモロガス座位（座位１０７２、１０７３、１０７４、および１０７５）を伴って表示されている。

【００９１】マルチプル生体ビューア１０５２は、本発明のグラフィカルビューアによる生体分子配列情報の効率的な伝送を示す別の例である。上述したように、太字のイタリック体かまたは特定の色を使用する等の他の方法で、選択座位ならびにそのホモログ座位およびパラログ座位を表示することによって、対象座位として区別している。図１０Ｄに示されるように、グラフィック表示されたコンティグの対象座位がページ１０５２に配置されているため、種々のコンティグ上の近くの座位を、視覚的に容易に比較することができる。この視覚的な表示は、図１０Ｅを参照にしつつ以下で説明する補体機能を使用することにより、さらに向上させることが可能である。

【００９２】本発明のこの実施形態におけるそのようなグラフィカルビューアの機能を利用するには、マルチプル生体ビューアページ１０５０に設けられたプルダウンメニュー選択１０５３をクリックすれば良い。このメニュー選択は、図４Ａで説明したような機能を提供する「Ｆｉｌｅ」、「Ｖｉｅｗ」、「Ｈｅｌｐ」選択を含んでいる。「Ｓｈｏｗ」選択１０５４からは、図１０Ｃに示すウィンドウに一覧表示されてマルチプル生体ビューアにロードされた全座位のリストにアクセスすることができる。「Ｓｈｏｗ」プルダウンメニューから座位を１つ選択することにより、ユーザは、その座位を、それが属するコンティグとともに表示するか隠すかを決定することができる。座位をクリックすることにより、ユーザは、その座位を表示するか隠すかを決定することができる。「Ｓｈｏｗ」メニューを使用して、コンティグに関して同様な決定を行うことも可能である。

【００９３】補体メニュー選択１０５５により、ユーザは、データからの目立った情報の抽出を容易にするために、コンティグおよび座位のグラフィック表示を操作することができる。特に、補体メニュー選択１０５５により、ユーザは、マルチプル生体ビューア１０５２に表示された任意のコンティグ上で、逆補体を実施することが可能となる。こうすれば、ビューア１０５２に表示されたホモロガス座位およびパラロガス座位が同じ読み方向で表示されるため、ユーザは、対象座位の近くの関連座位のパターンをより容易に見つけることができる。図１０Ｅに示されるマルチプル生体ビューアページ１０８０では、ページ１０５０に示された対象座位を、補体機能の使用によって同じ読み方向で示すことにより、コンティグ１０６２，１０６３，１０６５の逆補体を表示している。

【００９４】本明細書で説明する他の機能と同様、この補体機能へのショートカットも、当業者に周知の方法によって使用可能とすることができる。例えば、座位（コンティグ）の補体を表示するには、グラフィカルビューアが作動しているコンピュータシステムと接続されたキーボード上のシフトキーを押さえながら、そのコンティグをクリックすれば良い。

【００９５】図１１は、本発明の好ましい一実施形態におけるマルチプル生体ビューアの操作を一般化したプロ

セスを示すフローチャートである。プロセス１１００はステップ１１０１でスタートし、ステップ１１０２では、マルチプル生体ビューアシステムが、例えば図１０Ａのターゲットビューア内の座位のクリックに伴い、対象座位の選択を受け入れる。ステップ１１０４では、選択された対象座位に対するホモロガス座位またはパラロガス座位を含むライブラリのリストが、ウィンドウに表示される。この表示は、例えば図１０Ａのマルチプル生体ボタン等のボタンをユーザがクリックすることによって始まる。次に、ステップ１１０６において、システムは、リストから１つまたはそれ以上のライブラリの選択を受け取る。ステップ１１０８では、選択されたライブラリから得られるリスト、すなわち、選択された対象座位に対するホモロガス座位またはパラロガス座位のリストが、ウィンドウに表示される。ステップ１１１０において、システムは、表示されたリストから、選択された対象座位に対するホモロガス座位またはパラロガス座位の選択を受け取る。次いでステップ１１１２において、選択座位およびその関連のコンティグがマルチプル生体グラフィカルビューアにグラフィック表示される。好ましい一実施形態において、ビューアは、グラフィック表示されたデータの比較を容易にするために、全てのコンティグおよび座位を１つのパネル内に表示する。プロセスはステップ１１１４で終了する。

【００９６】実施：本発明は、システムまたは方法として実施でき、さらに、本明細書に記述した種々の操作を実行するためのプログラム命令等を含んだ種々のコンピュータ読み取り可能な媒体に組み込むことができる。上述したように、上記のシステムは、生体分子配列リレーショナルデータベースシステムとの関連のもとで実施されることが好ましい。上記の方法は、コンピュータで実行される方法であり、一般に、そのようなシステムの操作を含む。上記の媒体は、任意のコンピュータ読み取り可能な媒体で良い。コンピュータ読取り可能な媒体の例としては、ハードディスク、フロッピィディスク、磁気テープ等の磁気媒体、ＣＤ－ＲＯＭディスク等の光媒体、フロプティカルディスク等の光磁気媒体、ならびに読み取り専用メモリ（ＲＯＭ）およびランダムアクセスメモリ（ＲＡＭ）等の、プログラム命令を格納および実行するために構成されたハードウェアデバイス等が挙げられるが、これらに限定はされない。本発明はまた、エアウェーブ、光路、電線路等の適切な媒体を通じて伝わるような、搬送波に組み込まれていても良い。

【００９７】結論：以上では、理解を明確にする目的で本発明を細部にわたって説明したが、添付した特許請求の範囲の範囲内であれば、一定の変更および修正を加えられることは明らかである。また、本発明による方法、媒体、およびシステムを実施する代替の方法が、数多く存在することを留意する必要がある。上述したように、本発明の範囲は、本発明との関連のもとで説明した、細菌ゲノムデータベースシステムにおける使用に限定されるものではない。当業者ならば、種々のコンピュータベースの生体分子配列データベースシステムとの関連のもとで本発明を使用する方法を、本明細書でなされた説明から理解することが可能である。例えば、本発明のグラフィカルビューアを、他のタイプおよび形式の核酸配列の格納および分析や、発現核酸もしくはアミノ酸の配列の格納および分析に利用されるデータベースシステムとの関連のもとで使用することも可能である。本実施形態は例示であり、また、非限定的なものであり、本発明は、本明細書で述べた細目に限定されず、添付した特許請求の範囲の範囲および均等物の範囲内で変更することが可能である。

【図面の簡単な説明】

【図１Ａ】本発明の一実施形態に従ってデータベースサービスを提供するためのクライアントサーバイントラネットを示すブロック図である。

【図１Ｂ】ユーザからの照会に応えて生物学的情報を提供するために、図１Ａのクライアントサーバイントラネットで利用される種々のソフトウェア文書およびエンティティを示す概略図である。

【図１Ｃ】本発明の好ましい一実施形態に従ったグラフィカルビューア機能のアクセス可能性を、生体分子配列データベースに関連させて示すブロック図である。

【図２】本発明の一実施形態に従った生体分子配列グラフィカルビューアで表示される座位の選択に適した、ゲノム配列データベースのグラフィカルユーザインタフェースとしてのコンティグ結果ページを示す画面（ＨＴＭＬページ）である。

【図３】本発明の一実施形態に従った生体分子配列グラフィカルビューアにアクセスするのに適した、ゲノム配列データベースのグラフィカルユーザインタフェースとしての座位情報ページを示す画面である。

【図４Ａ】本発明の一実施形態に従った生体分子配列グラフィカルビューアのメインページを示す画面である。

【図４Ｂ】ズーム機能を示すために本発明の一実施形態に従って変更を加えられた生体分子配列グラフィカルビューアのメインページを示す画面である。

【図５Ａ】本発明の一実施形態に従った選択オブジェクト詳細ウィンドウである。

【図５Ｂ】本発明の一実施形態に従ったアライメントビューアを示す画面である。

【図５Ｃ】本発明の一実施形態に従ったアライメントビューアを示す画面である。

【図６】新しい開始座標機能を示すために本発明の一実施形態に従って変更を加えられた生体分子配列グラフィカルビューアのメインページを示した画面である。

【図７】新しいコンティグＩＤ機能を示すために本発明の一実施形態に従って変更を加えられた生体分子配列グラフィカルビューアのメインページを示した画面であ

る。

【図８Ａ】実際の生体分子配列を表示する機能を本発明の一実施形態に従って示した生体分子配列グラフィカルビューアのページを示す画面である。

【図８Ｂ】実際の生体分子配列を表示する機能を本発明の一実施形態に従って示した生体分子配列グラフィカルビューアのページを示す画面である。

【図８Ｃ】実際の生体分子配列を表示する機能を本発明の一実施形態に従って示した生体分子配列グラフィカルビューアのページを示す画面である。

【図９】本発明の一実施形態に従った生体分子配列グラフィカルビューアで遺伝子座位情報を見るためのプロセスフローを示したフローチャートである。

【図１０Ａ】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

【図１０Ｂ】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

【図１０Ｃ】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

【図１０Ｄ】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

【図１０Ｅ】本発明の一実施形態に従ったマルチプル生体分子配列グラフィカルビューアの操作を示す画面である。

【図１１】本発明の一実施形態に従った生体分子配列グラフィカルビューアでマルチプル生体の遺伝子座位情報を見るためのプロセスフローを示したフローチャートである。

【図１Ａ】

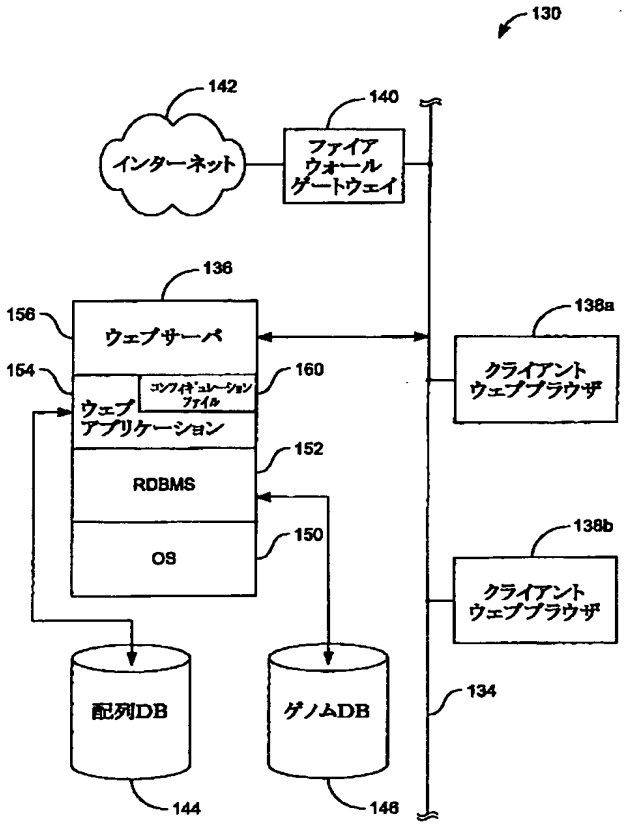


FIG. 1A

【図１Ｂ】

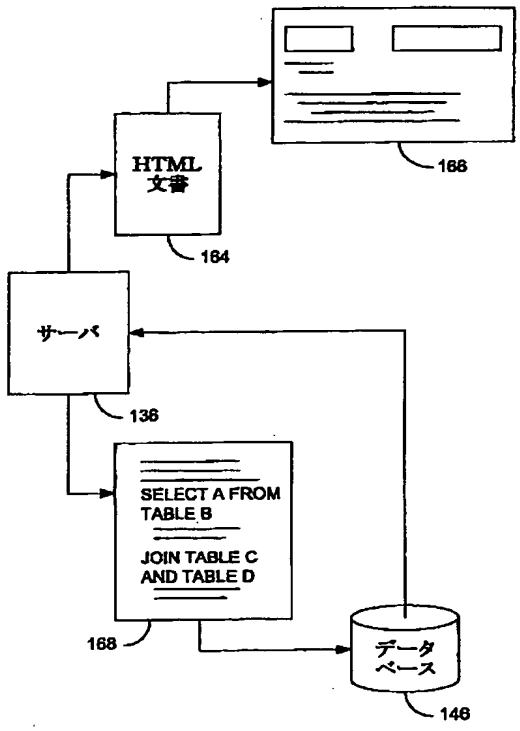


FIG. 1B

【図１Ｃ】

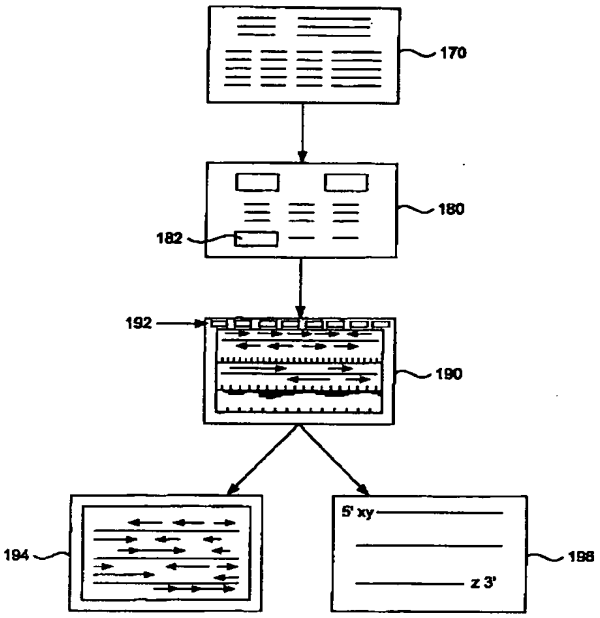


FIG. 1C

【図８Ｂ】

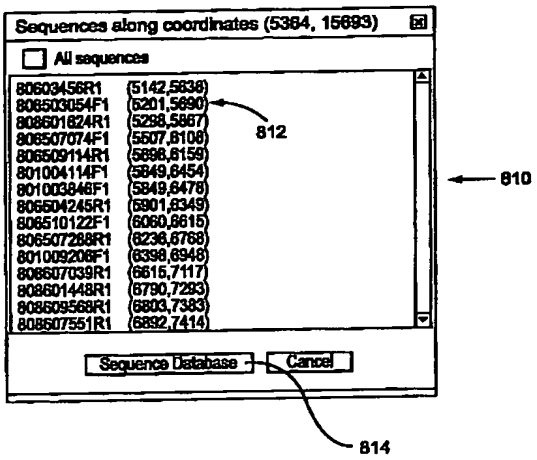


FIG. 8B

【図１０Ｂ】

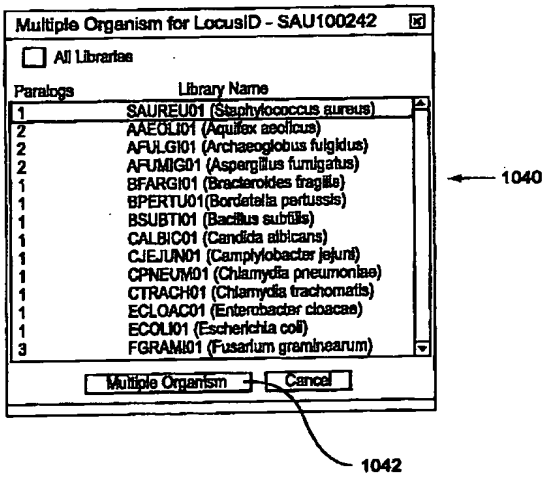


FIG. 10B

【図１０Ｃ】

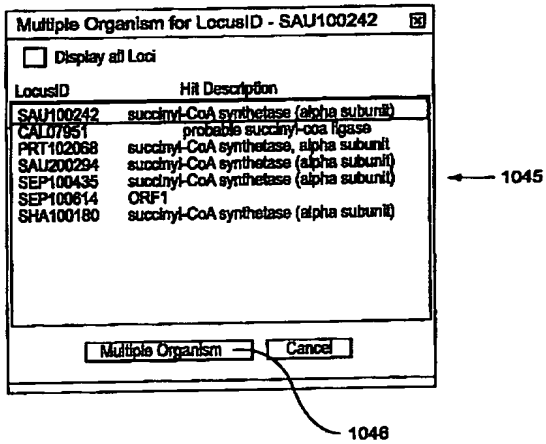


FIG. 10C

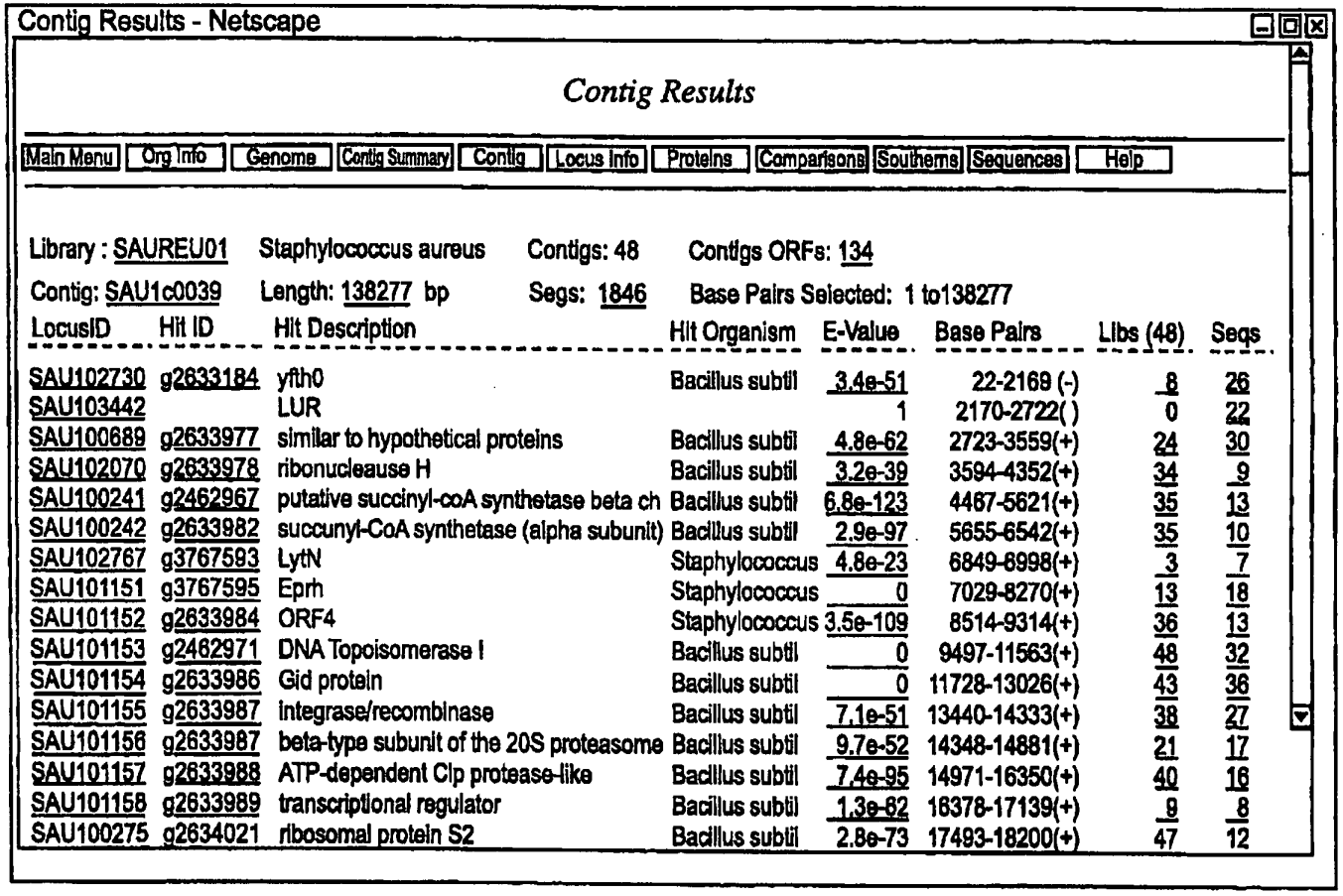


FIG. 2

200

【図２】

特開２００１－１２５９２９

【図３】

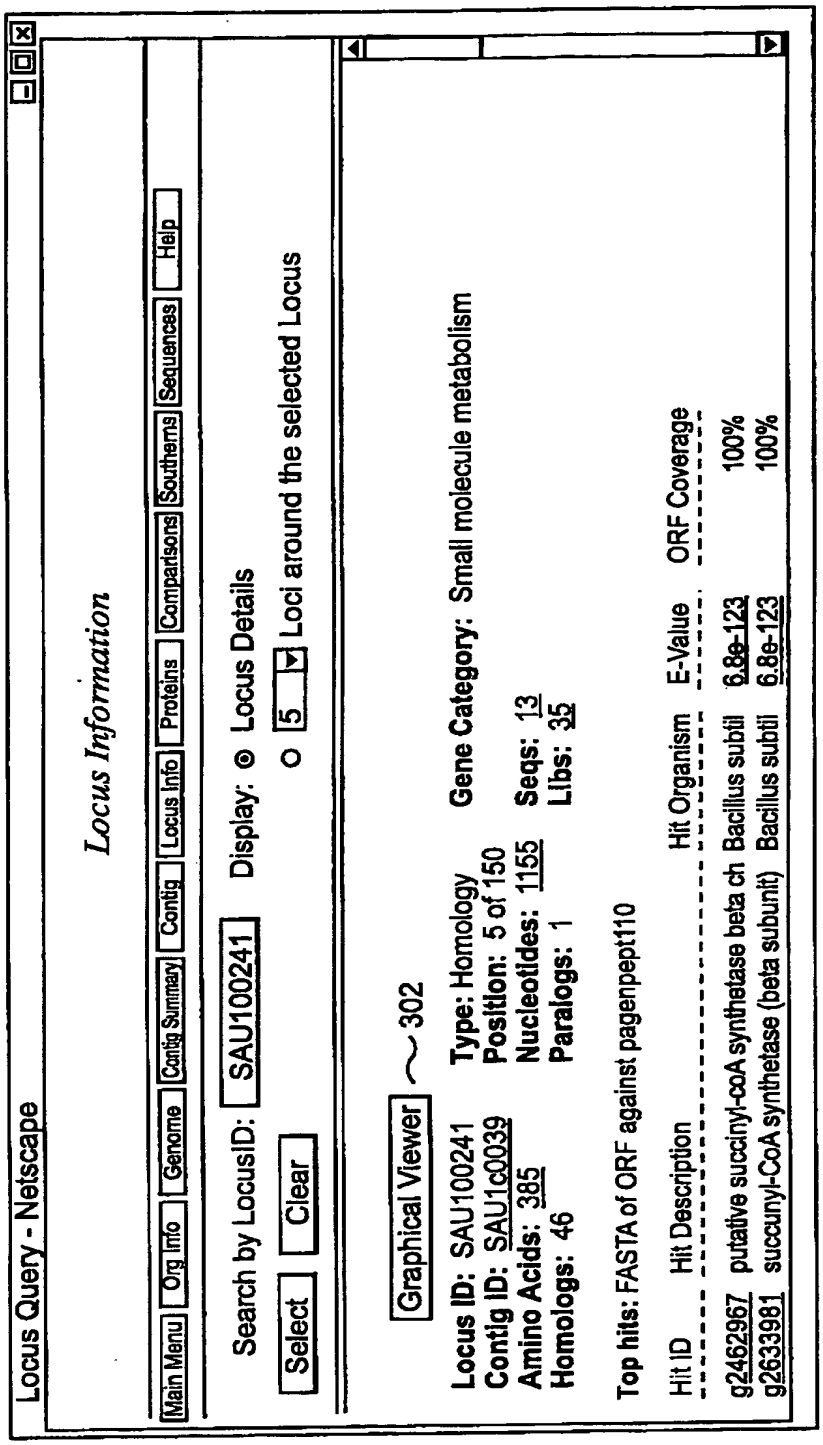

Locus Query - Netscape
Locus Information
Main Menu
Org Info
Genome
Contig Summary
Contig
Locus Info
Proteins
Comparisons
Southerns
Sequences
Help
Search by LocusID: SAU100241
Display: Locus Details
Select
Clear
5
Loci around the selected Locus
Graphical Viewer
302
Locus ID: SAU100241
Contig ID: SAU1c0039
Amino Acids: 385
Homologs: 46
Type: Homology
Position: 5 of 150
Nucleotides: 1155
Paralogs: 1
Gene Category: Small molecule metabolism
Seqs: 13
Libs: 35
Top hits: FASTA of ORF against pagenpept110
Hit ID
Hit Description
Hit Organism
E-Value
ORF Coverage
g2462967 putative succinyl-coA synthetase beta ch Bacillus subtil 6.8e-123 100%
g2633981 succunyl-CoA synthetase (beta subunit) Bacillus subtil 6.8e-123 100%
300

FIG. 3

【図４Ａ】

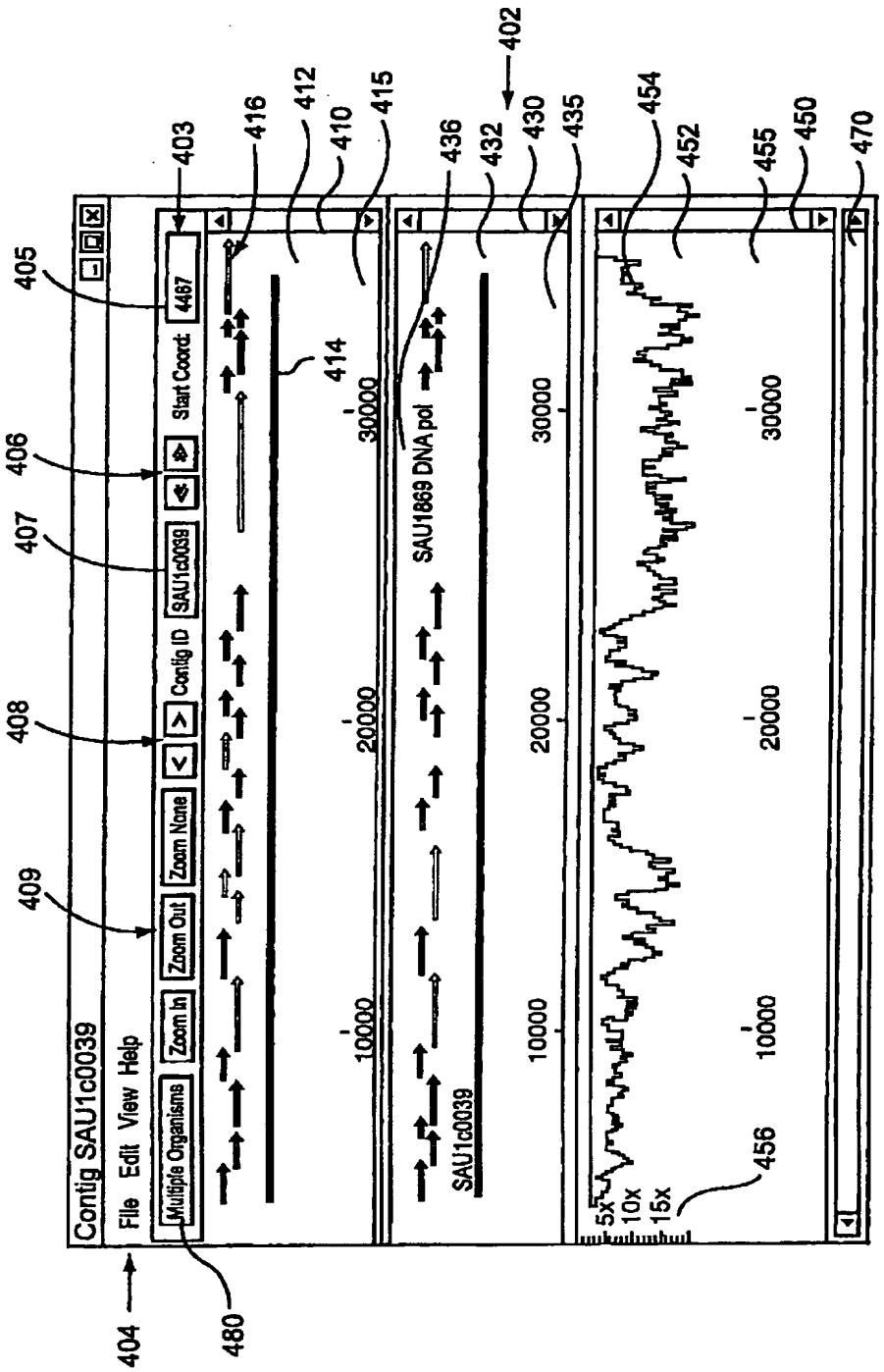

Contig SAU1c0039
File Edit View Help
Multiple Organisms
Zoom In
Zoom Out
Zoom None
Contig ID
SAU1c0039
Start Coord:
4487
SAU1869 DNA pol
SAU1c0039
5x
10x
15x
10000
20000
30000
400
402
403
404
405
406
407
408
409
410
412
414
415
416
430
432
435
436
450
452
454
455
456
470
480

FIG. 4A

【図４Ｂ】

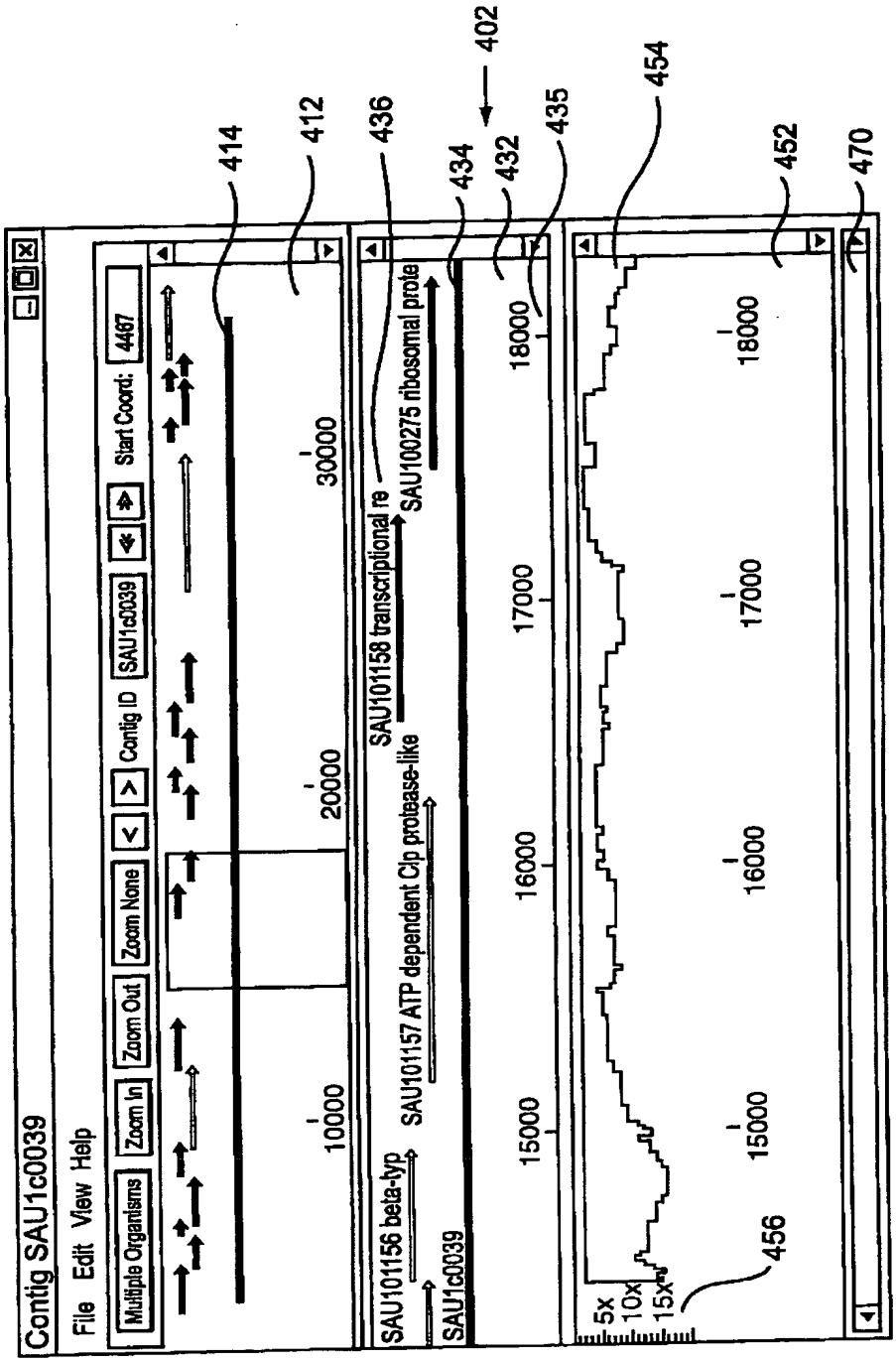

Contig SAU1c0039
File Edit View Help
Multiple Organisms
Zoom In
Zoom Out
Zoom None
Contig ID
SAU1c0039
Start Coord: 4467
414
412
436
434
432
435
454
452
470
402
10000
20000
30000
15000
16000
17000
18000
SAU101156 beta-typ
SAU101157 ATP dependent Clp protease-like
SAU101158 transcriptional re
SAU100275 ribosomal prote
SAU1c0039
5x
10x
15x
456

FIG. 4B

【図５Ａ】

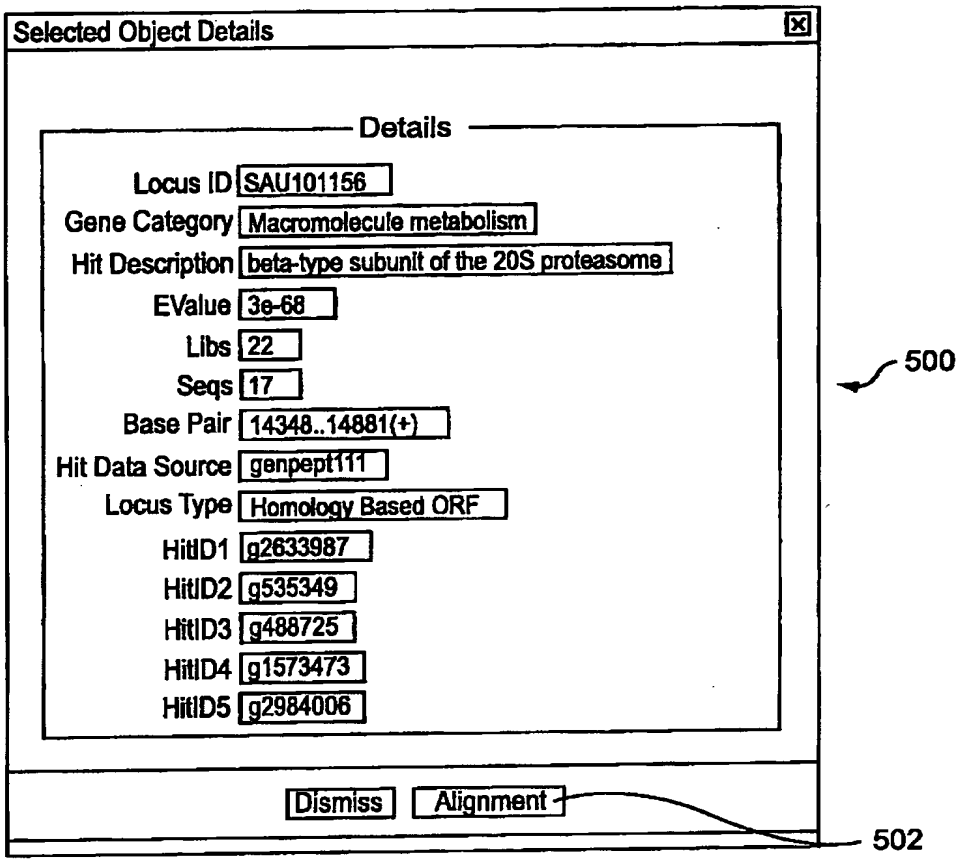

Selected Object Details
Details
Locus ID SAU101156
Gene Category Macromolecule metabolism
Hit Description beta-type subunit of the 20S proteasome
EValue 3e-68
Libs 22
Seqs 17
Base Pair 14348..14881(+)
Hit Data Source genpept111
Locus Type Homology Based ORF
HitID1 g2633987
HitID2 g535349
HitID3 g488725
HitID4 g1573473
HitID5 g2984006
Dismiss
Alignment
500
502

FIG. 5A

【図５Ｂ】

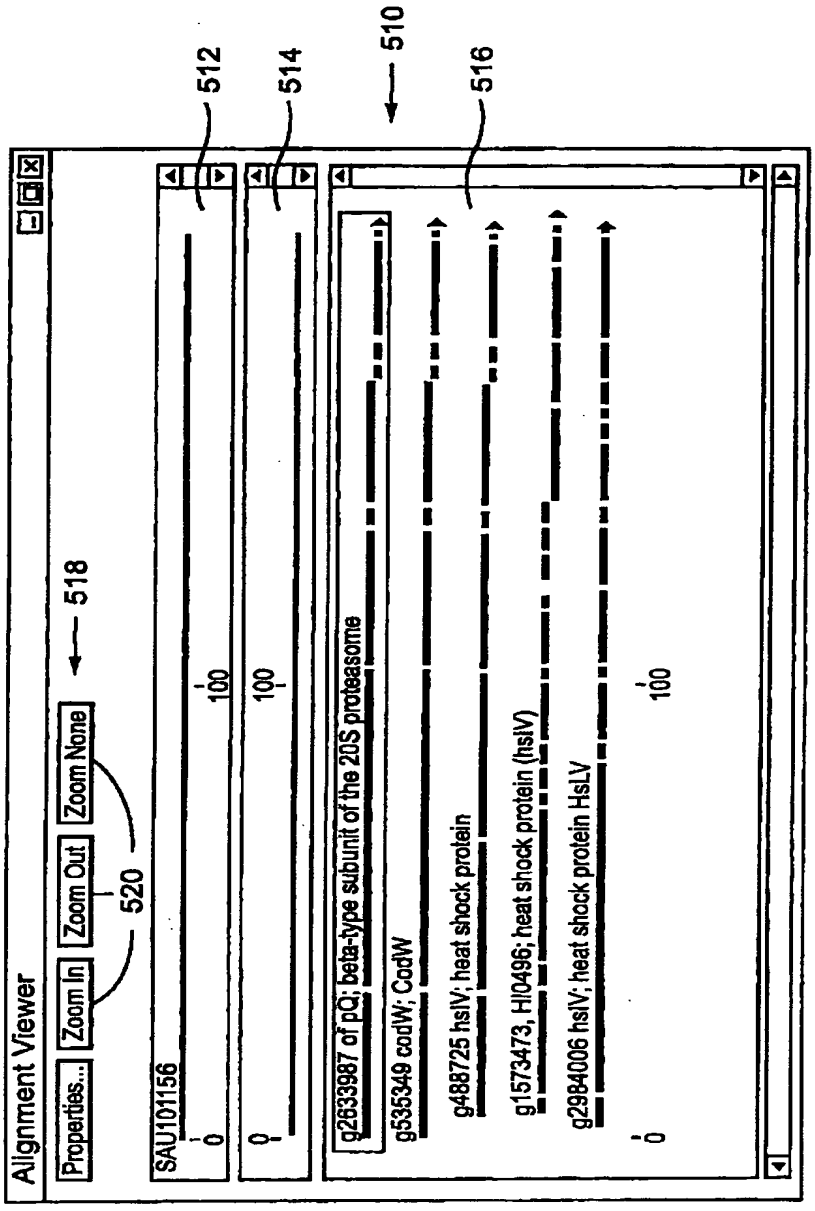

Alignment Viewer
Properties...
Zoom In
Zoom Out
Zoom None
518
520
512
514
510
516
SAU101156
0
100
0
100
0
100
g2833987 of pQ; beta-type subunit of the 20S proteasome
g535349 codW; CodW
g488725 hslV; heat shock protein
g1573473, HI0496; heat shock protein (hslV)
g2984006 hslV; heat shock protein HsLV

FIG. 5B

【図５Ｃ】

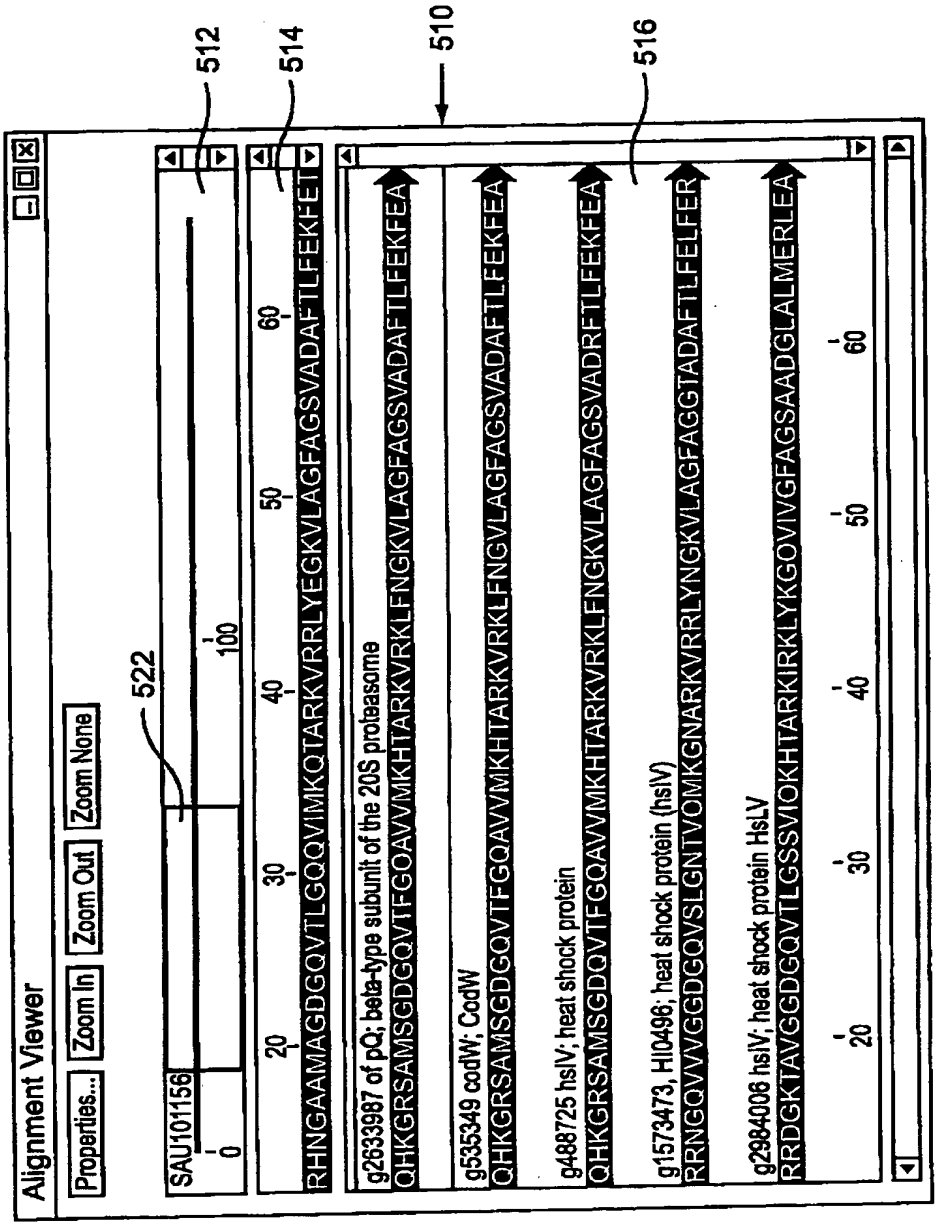

Alignment Viewer
Properties...
Zoom In
Zoom Out
Zoom None
SAU101156
0
100
522
512
514
510
516
20
30
40
50
60
RHNGAAMAGDGQVTLGQQVIMKQTARKVRRLYEGKVLAGFAGSVADAFTLFEKFET
g2633987 of pQ; beta-type subunit of the 20S proteasome
QHKGRSAMSGDGQVTFGQAVVMKHTARKVRKLFNGKVLAGFAGSVADAFTLFEKFEA
g535349 codW; CodW
QHKGRSAMSGDGQVTFGQAVVMKHTARKVRKLFNGVLAGFAGSVADAFTLFEKFEA
g488725 hslV; heat shock protein
QHKGRSAMSGDQVTFGQAVVMKHTARKVRKLFNGKVLAGFAGSVADRFTLFEKFEA
g1573473, HI0496; heat shock protein (hslV)
RRNGQVVVGGDGQVSLGNTVQMKGNARKVRRLYNGKVLAGFAGGTADAFTLFELFER
g2984006 hslV; heat shock protein HslV
RRDGKTAVGGDGQVTLGSSVIQKHTARKIRKLYKGQVIVGFAGSAADGLALMERLEA

FIG. 5C

【図６】

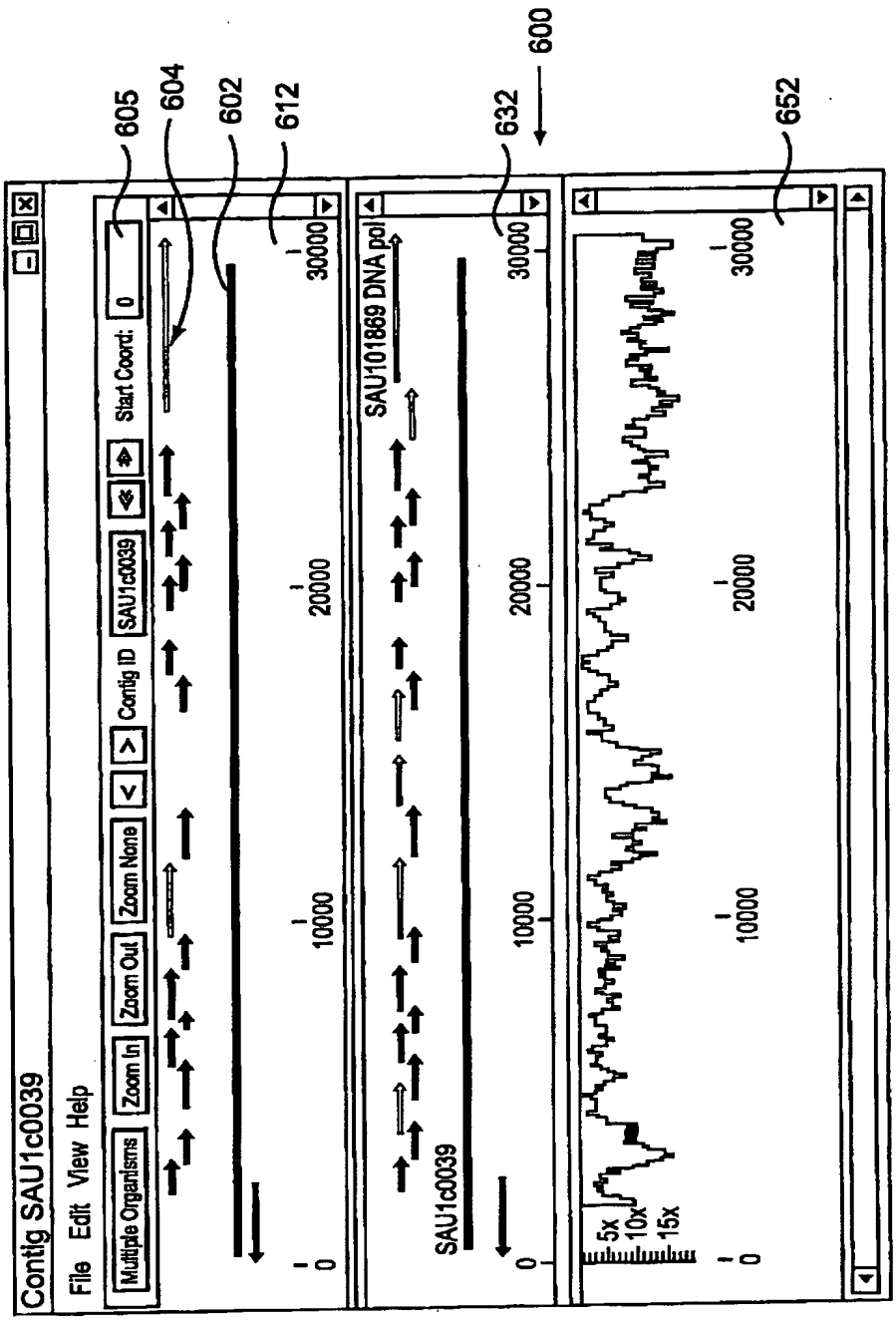

600
605
604
602
612
632
652
Contig SAU1c0039
File Edit View Help
Multiple Organisms
Zoom In
Zoom Out
Zoom None
Contig ID
SAU1c0039
Start Coord:
0
SAU101869 DNA pol
SAU1c0039
5x
10x
15x
0
10000
20000
30000

FIG. 6

【図７】

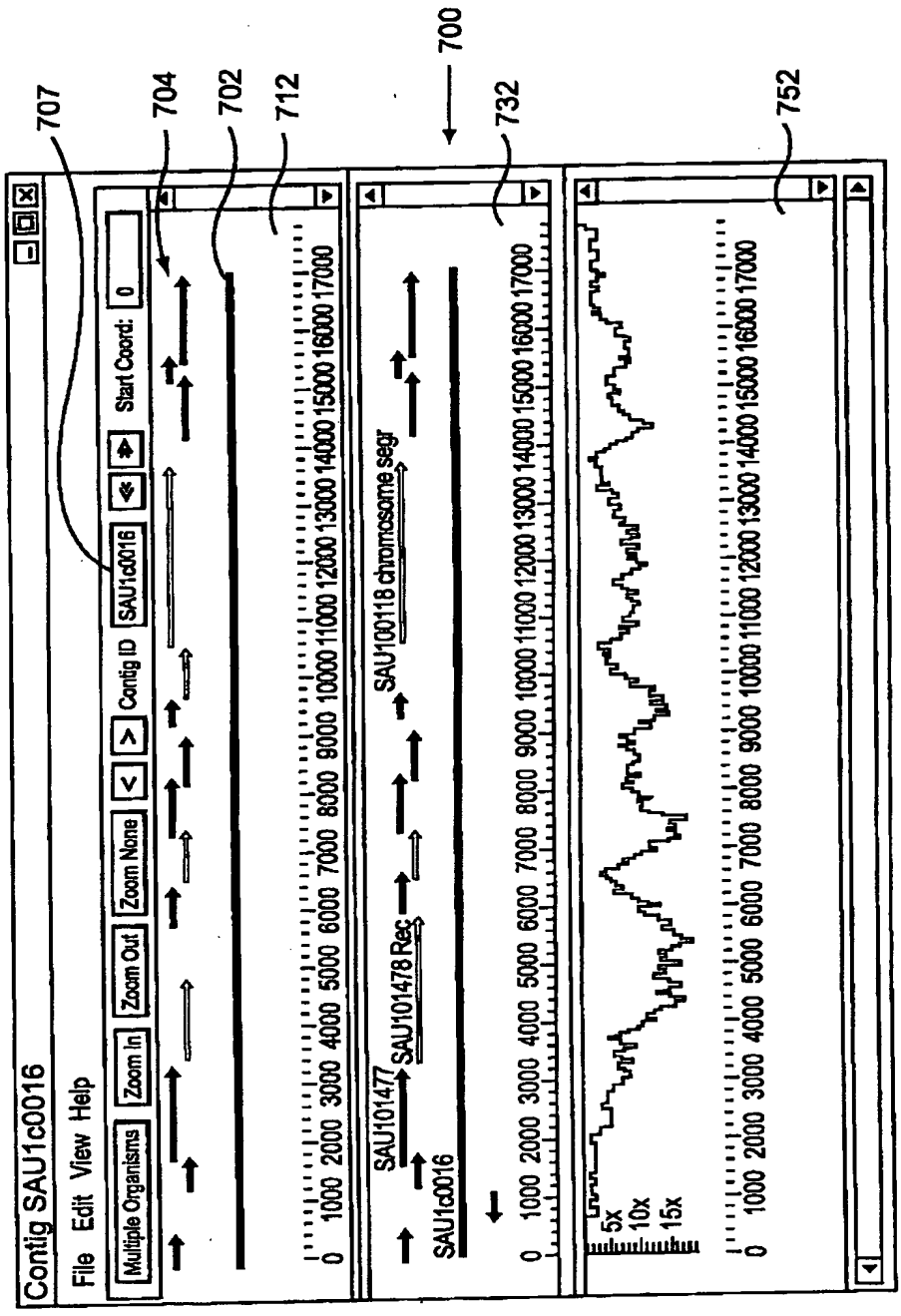

700
707
704
702
712
732
752
Contig SAU1c0016
File Edit View Help
Multiple Organisms
Zoom In
Zoom Out
Zoom None
Contig ID
SAU1c0016
Start Coord:
0
SAU101477
SAU101478 Rec
SAU100118 chromosome segr
SAU1c0016
5x
10x
15x
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 14000 15000 16000 17000

FIG. 7

【図８Ａ】

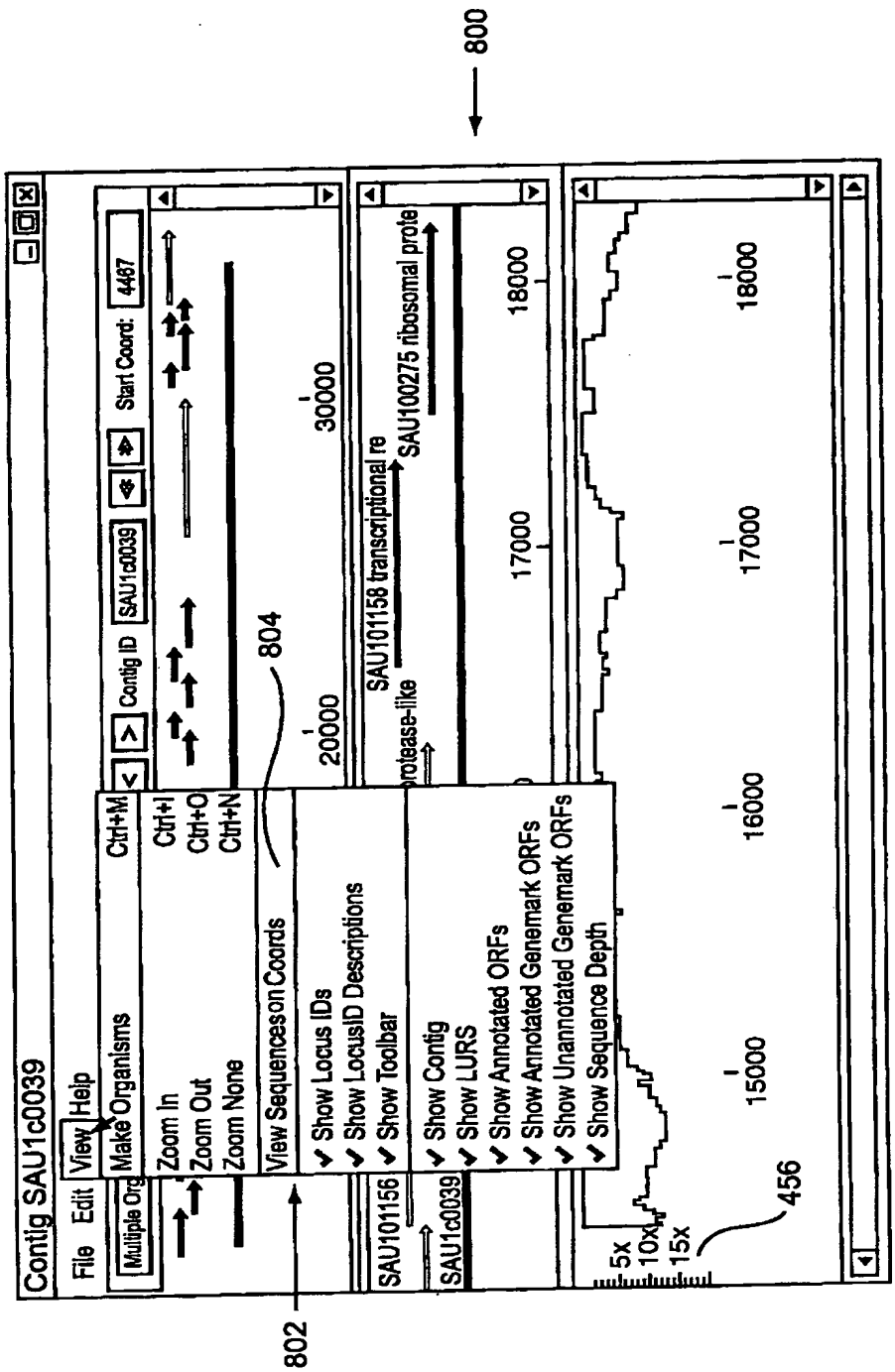

800
Contig SAU1c0039
File Edit View Help
Make Organisms Ctrl+M
Multiple Org
Zoom In Ctrl+I
Zoom Out Ctrl+O
Zoom None Ctrl+N
View Sequences on Coords Ctrl+N
Contig ID SAU1c0039
Start Coord: 4467
804
20000
30000
17000
18000
SAU101158 transcriptional re
SAU100275 ribosomal prote
protease-like
Show Locus IDs
Show LocusID Descriptions
Show Toolbar
Show Contig
Show LURS
Show Annotated ORFs
Show Annotated Genemark ORFs
Show Unannotated Genemark ORFs
Show Sequence Depth
SAU101156
SAU1c0039
5x
10x
15x
15000
16000
456
802

FIG. 8A

【図８Ｃ】

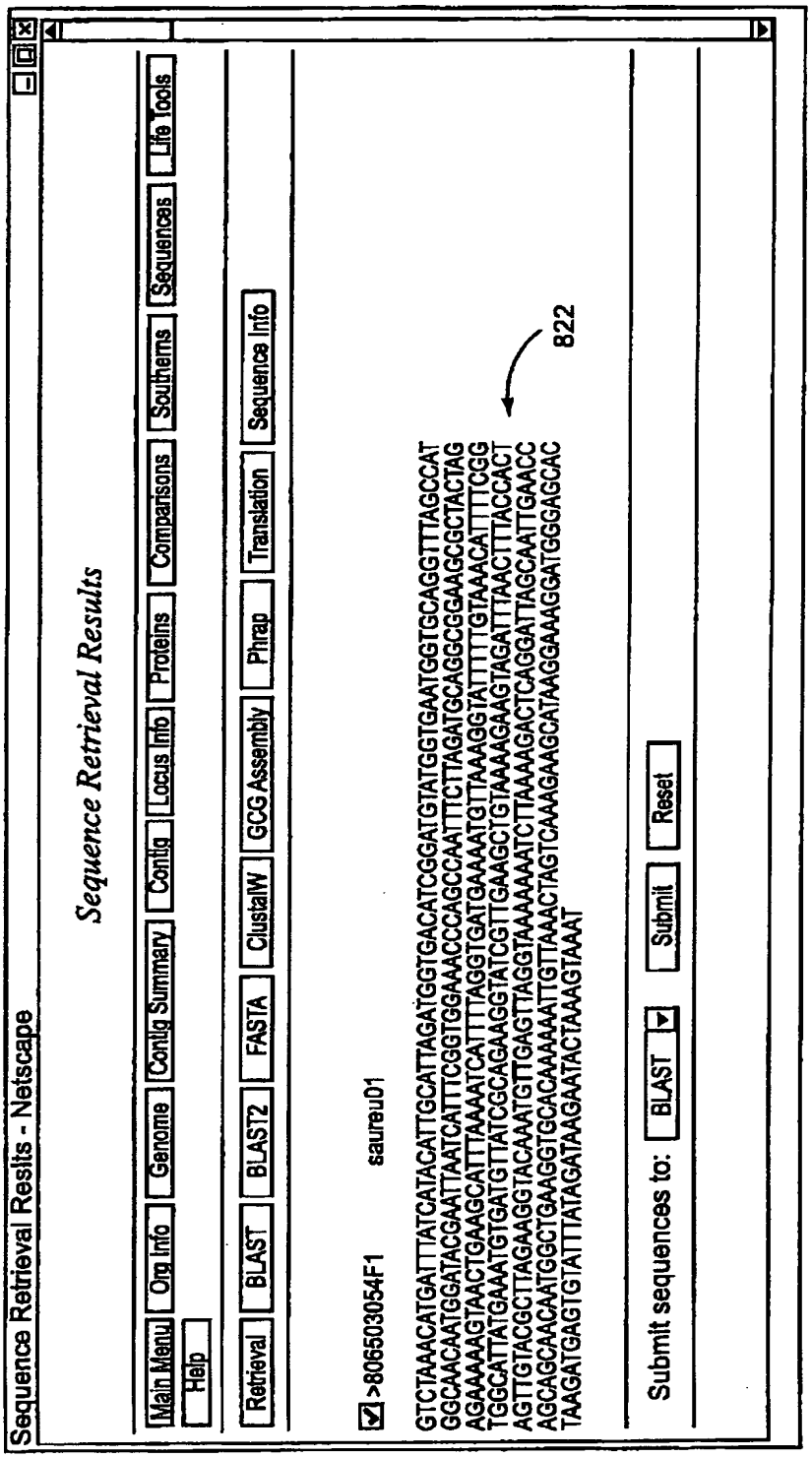

Sequence Retrieval Results - Netscape
Sequence Retrieval Results
Main Menu
Org Info
Genome
Contig Summary
Contig
Locus Info
Proteins
Comparisons
Southerns
Sequences
Life Tools
Help
Retrieval
BLAST
BLAST2
FASTA
ClustalW
GCG Assembly
Phrap
Translation
Sequence Info
>806503054F1 saureu01
GTCTAAACATGATTATCATACATTGCATTAGATGGTGACATCGGATGTATGGTGAATGGTGCAGGTTTAGCCAT
GGCAACAATGGATACGAATTAATCATTTCGGTGGAAACCCAGCCAATTTCTTAGATGCAGGCGGAAGCGCTACTAG
AGAAAAAGTAACTGAAGCATTTAAAATCATTTTAGGTGATGAAAATGTTAAAGGTATTTTTGTAAACATTTTCGG
TGGCATTATGAAATGTGATGTTATCGCAGAAGGTATCGTTGAAGCTGTAAAAGAAGTAGATTTAACTTTACCACT
AGTTGTACGCTTAGAAGGTACAAATGTTGAGTTAGGTAAAAAAATCTTAAAAGACTCAGGATTAGCAATTGAACC
AGCAGCAACAATGGCTGAAGGTGCACAAAAAATTGTTAAACTAGTCAAAGAAGCATAAGGAAAGGATGGGAGCAC
TAAGATGAGTGTATTTATAGATAAGAATACTAAAGTAAAT
822
Submit sequences to: BLAST
Submit
Reset
820

FIG. 8C

【図９】

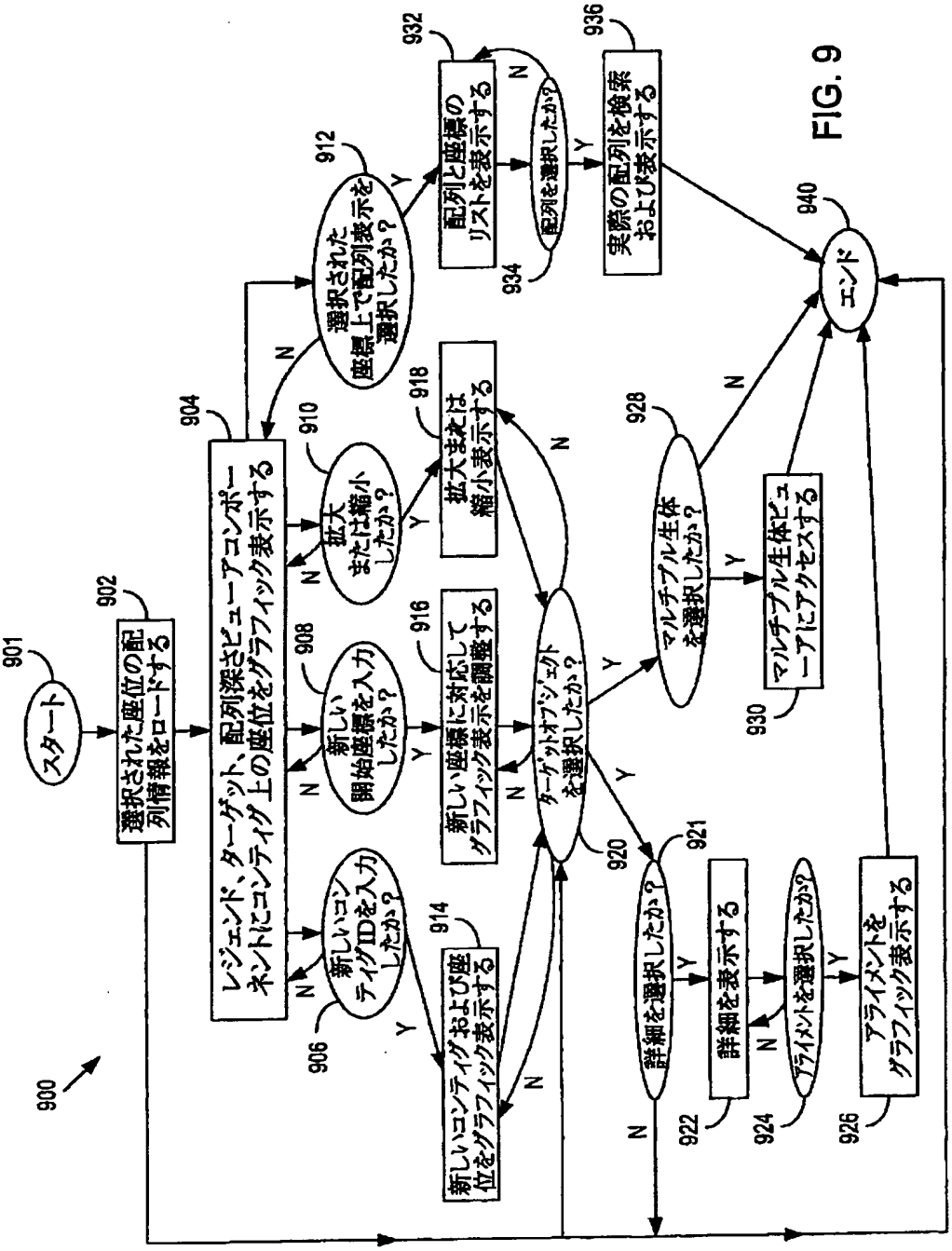

900
901
スタート
902
選択された座位の配列情報をロードする
904
レジェンド、ターゲット、配列深さビューアコンポーネントにコンティグ上の座位をグラフィック表示する
906
新しいコンティグIDを入力したか？
908
新しい開始座標を入力したか？
910
拡大または縮小したか？
912
選択された座標上で配列表示を選択したか？
914
新しいコンティグおよび座位をグラフィック表示する
916
新しい座標に対応してグラフィック表示を調整する
918
拡大または縮小表示する
920
ターゲットオブジェクトを選択したか？
921
詳細を選択したか？
922
詳細を表示する
924
アライメントを選択したか？
926
アライメントをグラフィック表示する
928
マルチプル生体を選択したか？
930
マルチプル生体ビューアにアクセスする
932
配列と座標のリストを表示する
934
配列を選択したか？
936
実際の配列を検索および表示する
940
エンド
Y
N

FIG. 9

【図１０Ａ】

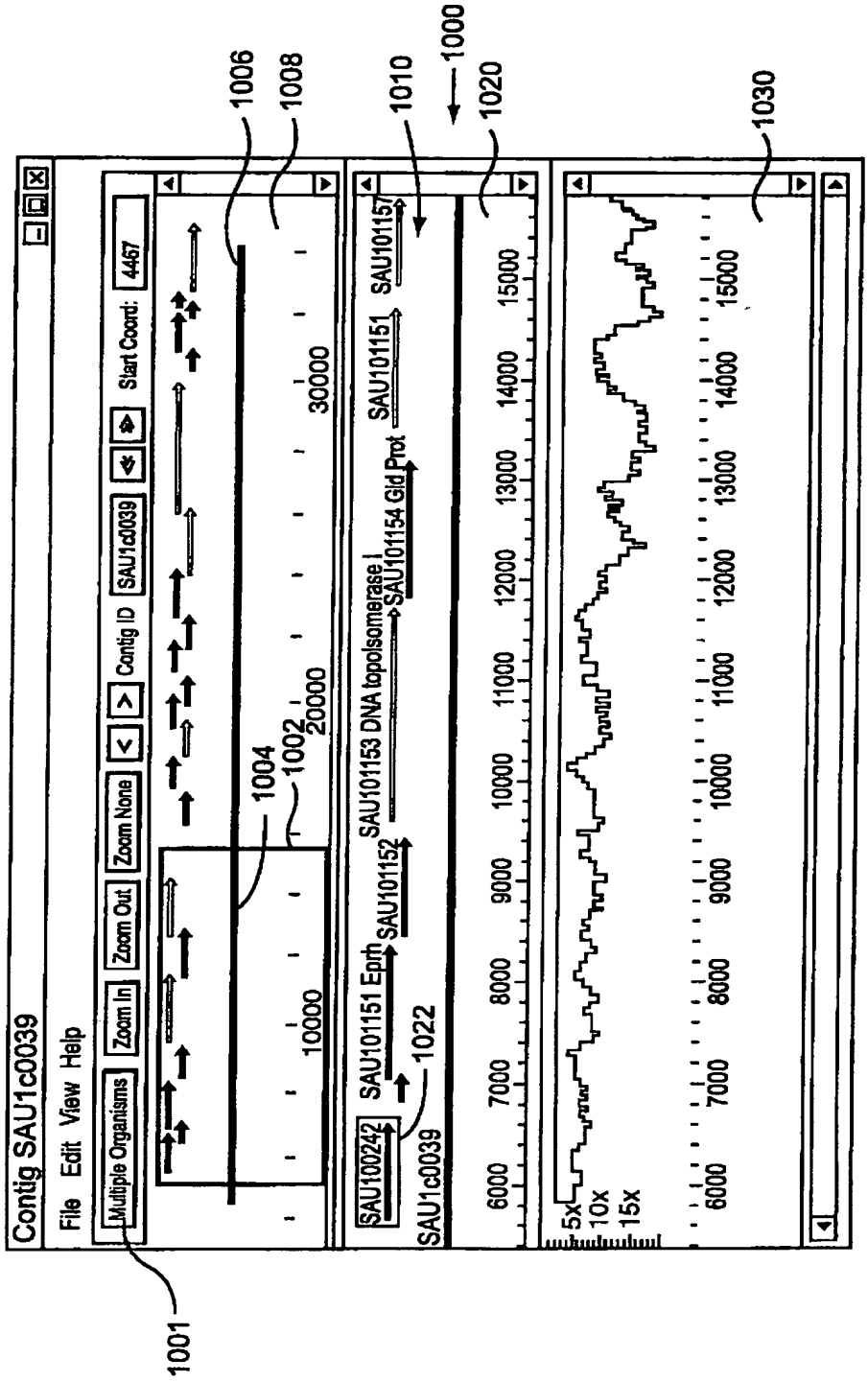

Contig SAU1c0039
File Edit View Help
Multiple Organisms
Zoom In
Zoom Out
Zoom None
Contig ID
SAU1c0039
Start Coord:
4467
1001
1002
1004
1006
1008
1010
1020
1022
1030
1000
SAU100242
SAU1c0039
SAU101151 Eprh
SAU101152
SAU101153 DNA topoisomerase I
SAU101154 Gid Prot
SAU101151
SAU101157
5x
10x
15x

FIG. 10A

【図１０Ｄ】

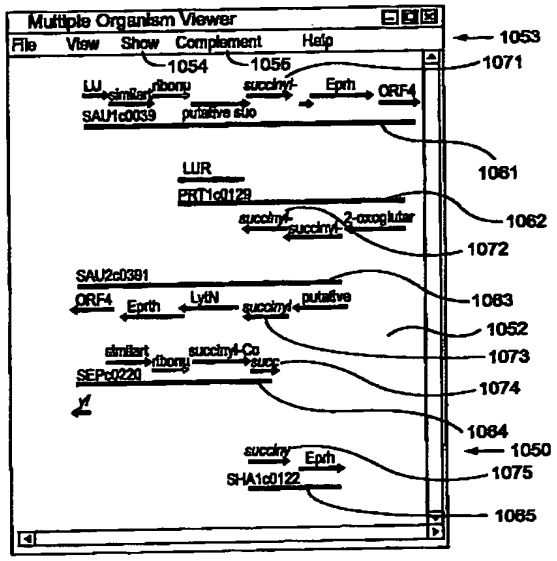


FIG. 10D

【図１０Ｅ】

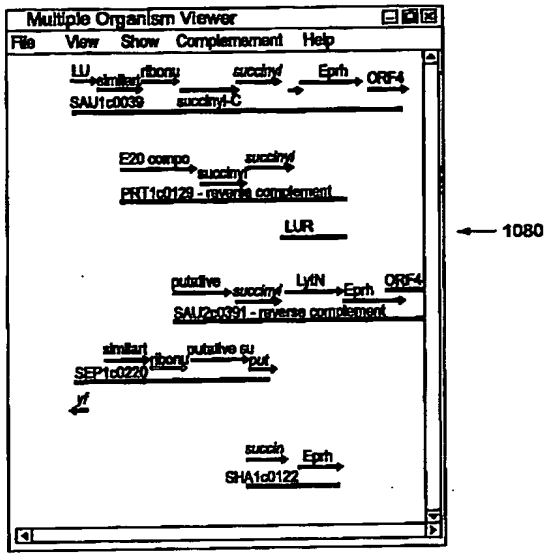


FIG. 10E

【図１１】

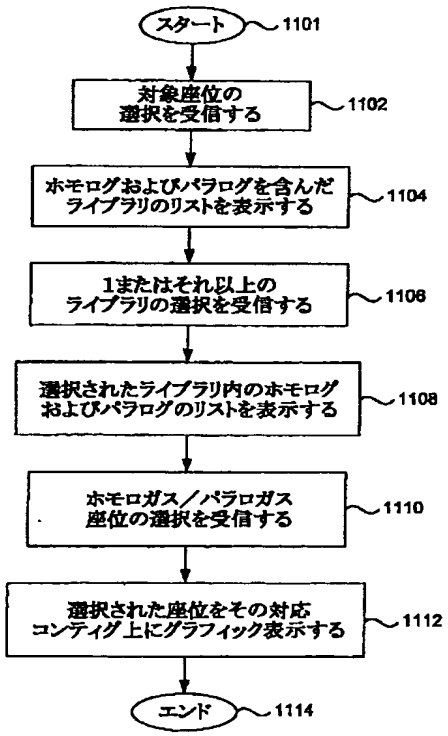


FIG. 11

フロントページの続き

(51)Int.Cl.[7]	識別記号	FI	ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　3/00	６５５	Ｇ０６Ｆ　3/00	６５５Ｂ
// Ｃ１２Ｎ　15/09		Ｃ１２Ｎ　15/00	Ａ

(72)発明者　フランク・ディー．・ルソー
アメリカ合衆国　カリフォルニア州94086　サニーベイル，ロゼッタ・コート，939
(72)発明者　ジョー・ドン・ヒース
アメリカ合衆国　カリフォルニア州95087　サニーベイル，グレープ・アベニュー，856
(72)発明者　ステファニー・エフ．・ベリー
アメリカ合衆国　カリフォルニア州95117　サン・ホセ，サニーゲート・コート，3564
(72)発明者　レイチェル・ジェイ．・ライト
アメリカ合衆国　カリフォルニア州94043　マウンテン・ビュー，アナ・アベニュー，339
(72)発明者　ピーター・エイ．・コヴィッツ
アメリカ合衆国　カリフォルニア州94116　サン・フランシスコ，29番・アベニュー，2442
(72)発明者　イヴォンヌ・アール．・ボルド
アメリカ合衆国　カリフォルニア州94087　サニーベイル，ジェイムスタウン・ドライブ，1109
(72)発明者　リー・クロフォード
アメリカ合衆国　カリフォルニア州94118　サン・フランシスコ　＃204，サクラメント・ストリート，3570

【外国語明細書】

1. TITLE OF INVENTION

Graphical Viewer for Biomolecular Sequence Data

2. CLAIMS

1. A method implemented in a computer system for presenting biomolecular sequence data, comprising:

retrieving biomolecular sequence data from a database in response to a user query; and

graphically depicting elements of the biomolecular sequence data in a user interface for said computer system.

2. The method of claim 1, wherein said graphical depiction comprises a plurality of panels, at least one of said plurality of panels graphically depicting depth of coverage information for a portion of a biomolecular sequence depicted in at least one other of said plurality of panels.

3. The method of claim 2, wherein said plurality of panels are comprised within a single frame.

4. The method of claim 3, wherein said plurality of panels provide graphical depictions representing different aspects of said biomolecular sequence data.

5. The method of claim 4, wherein said biomolecular sequence data comprises gene locus data.

6. The method of claim 5, wherein said plurality of panels comprises three panels.

7. The method of claim 6, wherein said three panels comprise a first panel graphically depicting at least a portion of a contig and its associated loci, a second panel graphically depicting at least a portion of the contig depicted in said first panel and annotated loci associated with the portion, and a third panel graphically depicting depth of coverage information for the portion of the contig depicted in the second panel.

8. The method of claim 7, wherein said third panel graphically depicts depth of coverage information for the portion of the contig depicted in the second panel.

9. The method of claim 1, wherein said method is implemented in Java programming language.

10. A method implemented in a computer system for presenting biomolecular sequence data, comprising:

retrieving biomolecular sequence data for a plurality of homologous loci from a database in response to a user query; and

graphically depicting at least some of the homologous loci in a user interface for said computer system.

11. The method of claim 10, wherein said graphical depiction comprises a single panel.

12. A computer system, comprising:

a database including biomolecular sequence data;

a user interface capable of

receiving a query relating to the biomolecular sequence data, and

graphically displaying the results of said query.

13. The system of claim 12, wherein said graphical depiction comprises a plurality of panels, at least one of said plurality of panels graphically depicting depth of coverage information for a portion of a biomolecular sequence depicted in at least one other of said plurality of panels.

14. The system of claim 13, wherein said plurality of panels are comprised within a single frame.

15. The system of claim 14, wherein said plurality of panels provide graphical depictions representing different aspects of said biomolecular sequence data.

16. The system of claim 15, wherein said biomolecular sequence data c

omprises gene locus data.

17. The system of claim 16, wherein said gene locus data is depicted in three panels comprising a first panel graphically depicting at least a portion of a contig and its associated loci, a second panel graphicall y depicting at least a portion of the contig depicted in said first panc l and annotated loci associated with the portion, and a third panel grap hically depicting depth of coverage information for the portion of the c ontig depicted in the second panel.

18. A computer-readable medium containing programmed instructions arr anged to graphically display biomolecular sequence data, the computer-re adable medium including programmed instructions for:

retrieving biomolecular sequence data from a computer system database in response to a user query; and

graphically depicting elements of the biomolecular sequence data in a user interface for the computer system.

3. DETAILED DESCRIPTION OF INVENTION

BACKGROUND OF THE INVENTION

The present invention relates generally to the field of bioinformatics.

In particular, the invention relates to methods, media and systems fo r graphically displaying computer-based biomolecular sequence informatio n.

Informatics is the study and application of computer and statistical tec hniques to the management of information. Bioinformatics includes the d evelopment of methods to search computer databases of biomolecular seque nce information (e.g., nucleic acid and protein) quickly, to analyze and display biomolecular sequence information, and to predict protein seque nce, structure and function from DNA sequence data.

Increasingly, molecular biology is shifting from the laboratory bench

to the computer desktop. Today's researchers require advanced quantitat ive analyses, database comparisons, and computational algorithms to expl ore the relationships between sequence and phenotype. Thus, by all acco unts, researchers cannot and will not be able to avoid using computer re sources to explore gene sequencing, gene expression, and molecular struc ture.

One use of bioinformatics involves studying an organism's genome to de termine the sequence and placement of its genes and their relationship t o other sequences and genes within the genome or to genes in other organ isms. Such information is of significant interest in biomedical and pha rmaceutical research, for instance to assist in the evaluation of drug e fficacy and resistance. To make genomic information manipulation easy t o perform and understand, sophisticated computer database systems have b een developed. Incyte Pharmaceuticals, Inc. of Palo Alto, CA, has devel oped several such databases, including some in which genomic sequence da ta is electronically recorded and annotated with information available f rom public sequence databases. Examples of such public sequence databas es include GenBank (NCBI) and SWISSPROT. The resulting information is s tored in a relational database that may be employed to determine relatio nships between sequences and genes within and among genomes.

While genetic data processing and relational database systems such as those developed by Incyte Pharmaceuticals, Inc. provide great power and flexibility in analyzing genetic information, further improvements in th ese systems will help accelerate biological research for numerous applic ations.

One area of interest in this regard is the display of biomolecular seq uence information. As noted above, an important goal of genome research to determine the sequence and placement of a organism's genes and their relationship to other sequences and genes within the genome, to genes i

n other organisms, and to related protein sequences. The ability to cle arly and effectively display gene loci information for a given organism or organisms would greatly assist this task.

Accordingly, the development of a display tool which allows a user to cl early and effectively display gene loci information for a given organism or organisms and/or other biomolecular sequence information is desirabl e.

SUMMARY OF THE INVENTION

The present invention meets this need by providing methods, media and sy stems for graphically displaying computer-based biomolecular sequence in formation. Generally, biomolecular sequence information may be graphica lly depicted in a variety of different forms in accordance with the pres ent invention. The sequence information may be composed of nucleotide o r amino acid sequence information or both. The graphical depictions may be in several different formats providing different information relatin g to the sequences, and may be displayed in one or more screens of a com puter user interface.

A graphical viewer in accordance with the present invention preferably h as a plurality of panels, each panel displaying information about the bi omolecular sequence data of interest in a different way on a single scre en or page. For example, a first panel could show a graphical represent ation of the entire biomolecular sequence, or the portion of the sequenc e of interest, with the locations of particular subsequences of interest indicated. A second panel could show a more detailed graphical represe ntation of all or a selected portion of the sequence represented in the first window, allowing a user to focus on a particular subsequence of in terest. This second panel view could depict additional information, suc h as annotations, relating to the particular subsequences of interest.

A third panel could show information graphically representing the confid

ence level or origination, for example, of the biomolecular sequence dat a represented in one or more of the other panels. Additional panels on the same or additional screens could show, for example, the actual nucle otide or amino acid sequence of or relating to a selected subsequence of interest represented in one or more of the other panels, or other infor mation relating to the biomolecular sequence data.

In one preferred embodiment, a graphical viewer in accordance with the p resent invention provides a graphical representation of all or a selecte d portion of an organism's genome with its individual loci indicated. T he viewer allows the user to focus on a particular region or locus of in terest and have it also be graphically represented with additional infor mation, such as annotations. A graphical depiction of sequence coverage for the sequence regions represented in the viewer may also be provided .

The viewer may also provide for the display of related loci from other p ortions of the organism's genome (i.e., paralogs), and allows for the re trieval of information about the loci, such as actual nucleotide sequenc es or detailed annotations, from an associated relational database syste m. In addition, a graphical viewer in accordance with the present inven tion may provide for the graphical representation and comparison of mul tiple portions of the genome of one or more organisms based on a locus o f interest and its corresponding paralogs and homologs (related loci fro m another organism's genome).

A graphical viewer in accordance with a preferred embodiment of the pres ent invention preferably provides graphical representations of the genom ic data in a plurality of panels, each panel displaying information abou t the genomic data of interest in a different way. In a particularly pr eferred embodiment of the invention, the graphical viewer has three main panels on a single screen: a legend viewer, which shows the entire por

tion of the genome under consideration; a target viewer, which allows a user to focus ("zoom in") on areas of the genome portion of particular interest; and a sequence depth viewer, which contains graphical information illustrating the depth of coverage over the length of the genome portion under consideration.

In one aspect, the present invention provides a method implemented in a computer system for presenting biomolecular sequence data. The method involves retrieving biomolecular sequence data from a database in response to a user query, and graphically depicting elements of the biomolecular sequence data in a user interface for the computer system. The graphical depiction may include a plurality of panels representing different aspects of the biomolecular sequence data in a single frame.

In a preferred embodiment, the biomolecular sequence data my include gene locus data and be graphically depicted in three panels, the first panel graphically depicting at least a portion of a contig and its associated loci, the second panel graphically depicting at least a portion of the contig depicted in the first panel and annotated loci associated with the portion, and the third panel graphically depicting information indicating the number of sequencing operations conducted to determine the sequence data depicted in the second panel. The third panel may graphically depicts sequences used to assemble the portion of the contig depicted in the second panel, or depth of coverage information for the portion of the contig depicted in the second panel.

In another aspect, the invention provides another method implemented in a computer system for presenting biomolecular sequence data. The method involves retrieving biomolecular sequence data for a plurality of homologous loci from a database in response to a user query, and graphically depicting at least some of the homologous loci in a user interface for the computer system.

In yet another aspect, the invention provides a computer system. The computer system includes a database including biomolecular sequence data, and a user interface. The user interface is capable of receiving a query relating to the biomolecular sequence data, and graphically displaying the results of the query.

In still another aspect, the invention provides a computer-readable medium containing programmed instructions arranged to graphically display biomolecular sequence data. The computer-readable medium includes programmed instructions for retrieving biomolecular sequence data from a computer system database in response to a user query, and graphically depicting elements of the biomolecular sequence data in a user interface for the computer system.

These and other features and advantages of the present invention will be presented in more detail in the following specification of the invention and the accompanying figures which illustrate by way of example the principles of the invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Reference will now be made in detail to preferred embodiments of the invention. Examples of the preferred embodiments are illustrated in the accompanying drawings. While the invention will be described in conjunction with these preferred embodiments, it will be understood that it is not intended to limit the invention to one or more preferred embodiments. On the contrary, it is intended to cover alternatives, modifications, and equivalents as may be included within the spirit and scope of the invention as defined by the appended claims. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. The present invention may be practiced without some or all of these specific details. In other instances, well known process operations have not been described in detail in

order not to unnecessarily obscure the present invention.

Introduction

The present invention provides methods, media and systems for graphicall y displaying computer-based biomolecular sequence information. Generall y, biomolecular sequence information may be graphically depicted in a va riety of different forms in accordance with the present invention. The sequence information may be composed of nucleotide or amino acid sequenc e information or both. The graphical depictions may be in several diffe rent formats providing different information relating to the sequences, and may be displayed in one or more screens of a computer user interface .

A graphical viewer in accordance with the present invention preferably h as a plurality of panels, each panel displaying information about the bi omolecular sequence data of interest in a different way on a single scre en or page. For example, a first panel could show a graphical represent ation of the entire biomolecular sequence, or the portion of the sequenc e of interest, with the locations of particular subsequences of interest indicated. A second panel could show a more detailed graphical represe ntation of all or a selected portion of the sequence represented in the first window, allowing a user to focus on a particular subsequence of in terest. This second panel view could depict additional information, suc h as annotations, relating to the particular subsequences of interest. A third panel could show information graphically representing the confid ence level or origination, for example, of the biomolecular sequence dat a represented in one or more of the other panels. Additional panels on the same or additional screens could show, for example, the actual nucle otide or amino acid sequence of or relating to a selected subsequence of interest represented in one or more of the other panels, or other infor mation relating to the biomolecular sequence data.

In one preferred embodiment, a graphical viewer in accordance with the present invention provides a graphical representation of all or a selected portion of an organism's genome with its individual loci indicated. The viewer allows the user to focus on a particular region or locus of interest and have it also be graphically represented with additional information, such as annotations. A graphical depiction of sequence coverage for the sequence regions represented in the viewer may also be provided.

The viewer also may also provide for the display of related loci from other portions of the organism's genome (i.e., paralogs), and allows for the retrieval of information about the loci, such as actual nucleotide sequences or detailed annotations, from an associated relational database system. In addition, a graphical viewer in accordance with the present invention may provide for the graphical representation and comparison of multiple portions of the genome of one or more organisms based on a locus of interest and its corresponding paralogs and homologs (related loci from another organism's genome).

A graphical viewer in accordance with a preferred embodiment of the present invention preferably provides graphical representations of the genomic data in a plurality of panels, each panel displaying information about the genomic data of interest in a different way. In a particularly preferred embodiment of the invention, the graphical viewer has three main panels on a single screen: a legend viewer, which always shows the entire portion of the genome under consideration; a target viewer, which allows a user to focus ("zoom in") on areas of the genome portion of particular interest; and a sequence depth viewer, which contains graphical information illustrating the depth of coverage over the length of the genome portion under consideration.

Of course, as noted above, a graphical viewer in accordance with the pre

sent invention may be used to display biomolecular sequence information other than the gene locus information described with reference to the preferred embodiments of the invention described herein. For example, a graphical viewer in accordance with the present invention may be used to display peptide or nucleotide sequence information, and can be used to display actual sequences resulting from comparisons of sequences from, for example, a BLAST or FASTA search.

The Graphical Viewer Environment

As noted above, a graphical viewer in accordance with the present invention is preferably used in connection with a biomolecular sequence relational database system, such as those developed by Incyte Pharmaceuticals, Inc. of Palo Alto, CA, and described, for example, in United States Patent Nos. 5,970,500; 5,953,727; 5,966,712 and 6,023,659. Data to be displayed by a graphical viewer in accordance with the present invention is accessed from such a database system using techniques and commands well known to those of skill in the art. Figures 1A and 1B and the associated description provided below provide a context in which a graphical viewer in accordance with the present invention may operate.

Figure 1A depicts a network system 130 suitable for storing and retrieving information in relational databases, such as those suitable for supporting a graphical viewer in accordance with the present invention. Network 130 includes a network cable 134 to which a network server 136 and clients 138a and 138b (representative of possibly many more clients) are connected. Cable 134 is also connected to a firewall/gateway 140 which is in turn connected to the Internet 142.

Network 130 may be any one of a number of conventional network systems, including a local area network (LAN) or a wide area network (WAN), as is known in the art (e.g., using Ethernet, IBM Token Ring, or the like).

The network includes functionality for packaging client calls in a wel

l-known format (e.g., URL) together with any parameter information into a format (of one or more packets) suitable for transmission across a cable or wire 134, for delivery to database server 136.

Server 136 includes the hardware necessary for running software to (1) access database data for processing user requests, and (2) provide an interface for serving information to client machines 138a and 138b. In a preferred embodiment, depicted in Figure 1A, the software running on the server machine supports the World Wide Web protocol for providing page data between a server and client. In this embodiment, a web server 156 having URL and HTTP functionality communicates with a client via the HTTP protocol.

Client/server environments, database servers, relational databases and networks are well documented in the technical, trade, and patent literature. For a discussion of database servers, relational databases and client/server environments generally, and SQL servers particularly, see, e.g., Nath, A., The Guide To SQL Server, 2nd ed., Addison-Wesley Publishing Co., 1995.

As shown, server 136 includes an operating system 150 (e.g., UNIX) on which runs a relational database management system 152, a World Wide Web application 154, and a World Wide Web server 156. The software on server 136 may assume numerous configurations. For example, it may be provided on a single machine or distributed over multiple machines.

World Wide Web application 154 includes the executable code necessary for generation of database language statements (e.g., Standard Query Language (SQL) statements). Generally, the executables will include embedded SQL statements. In addition, application 154 includes a configuration file 160 which contains pointers and addresses to the various software entities that comprise the server as well as the various external and internal databases which must be accessed to service user requests. Conf

iguration file 160 also directs requests for server resources to the app ropriate hardware -- as may be necessary should the server be distribute d over two or more separate computers.

Each of clients 138a and 138b includes a World Wide Web browser for pr oviding a user interface to server 136, and including code necessary to generate HTML pages. Through the Web browser, clients 138a and 138b con struct search requests for retrieving data from a sequence database 144 and/or a genomic database 146, for example. Thus, the user will typical ly point and click to user interface elements such as buttons, pull dowo menus, scroll bars, etc. conventionally employed in graphical user inte rfaces. The requests so formulated with the client's Web browser are tr ansmitted to Web application 154 which formats them to produce a query t hat can be employed to extract the pertinent information from sequence d atabase 144 or genomic database 146.

In the embodiment shown, the Web application accesses data in genomic database 146 by first constructing a query in a database language (e.g., Sybase or Oracle SQL). The database language query is then handed to r elational database management system 152 which processes the query to ex tract the relevant information from database 146. In the case of a requ est to access sequence database 144, Web application 154 directly commun icates the request to that database without employing the services of da tabase management system 152.

The procedure by which user requests are serviced is further illustrat ed with reference to Figure 1B. In this embodiment, the World Wide Web server and/or executable Web application components of server 136 provid e Hypertext Mark-up Language documents ("HTML pages") 164 to a client ma chine. At the client machine, the HTML document provides a user interfa ce 166 which is employed by a user to formulate his or her requests for access to database 146. That request is converted by the Web applicatic

n component of server 136 to a SQL query 168. That query is used by the database management system component of server 136 to access the relevant data in database 146 and provide that data to server 136 in an appropriate format. Server 136 then generates a new HTML document, possibly through the Web application 154, relaying the database information to the client as a view in user interface 166.

While the embodiment shown in Figure 1A employs a World Wide Web server and World Wide Web browser for a communication between server 136 and clients 138a and 138b, other communications protocols will also be suitable. For example, client calls may be packaged directly as SQL statements, without reliance on Web application 154 for a conversion to SQL. Clients may also query the database directly without using a client browser.

When network 130 employs a World Wide Web server and clients, it must support a TCP/IP protocol. Local networks such as this are sometimes referred to as "Intranets." An advantage of such Intranets is that they allows easy communication with public domain databases residing on the World Wide Web (e.g., the GenBank World Wide Web site). Thus, in a particular preferred embodiment of the present invention, clients 138a and 138b can directly access data (via Hypertext links for example) residing on Internet databases using a HTML interface provided by Web browsers and Web server 156.

Bear in mind that if the contents of the local databases are to remain private, a firewall 140 must preserve in confidence the contents of a sequence database 144 and a genomic database 146.

In a preferred embodiment, sequence database 144 is a flat file database with a single file for genomic sequences from different species. Other possible approaches may include partitioning the sequence data according to different species or whether or not sequences have been found to

be unique to the local database (i.e., sequences that did not have any h its in an external database such as GenBank).

Preferably, the information in genomic database 146 is stored in a rel ational format. Such a relational database supports a set of operations defined by relational algebra. It generally includes tables composed o f columns and rows for the data contained in the database. Each table h as a primary key, being any column or set of columns the values of which uniquely identify the rows in the table. The tables of a relational da tabase may also include a foreign key, which is a column or set of colum ns the values of which match the primary key values of another table. A relational database is also generally subject to a set of operations (s elect, project, product, join and divide) which form the basis of the re lational algebra governing relations within the database. As noted abov e, relational databases are well known and documented (see, e.g., Nath, A., The Guide To SQL Serve, referenced above).

A relational database may be implemented in different ways. In Oracle (trademark) databases, for example, the various tables are not physicall y separated, as there is one instance of work space with different owner ship specified for different tables. In Sybase(trademark) databases, in contrast, the tables may be physically segregated into different "datab ases."

One specific configuration for network 130 for multiple users provides both the genomic and sequence databases on the same machine. If there is a high volume of sequence searching, it may be desirable to have a se cond processor of similar size and split the application across the two machines to improve response time.

A suitable dual processor server machine may be any of the following w orkstations: Sun - Ultra-Sparc 2(trademark) (Sun Microsystems, Inc. of M ountain View, CA), SGI - Challenge L(trademark) (Silicon Graphics, Inc.

of Mountain View, CA), and DEC - 2100A(trademark) (Digitial Electronics Corporation of Maynard, MA). Multiprocessor systems (minimum of 4 processors to start) may include the following: Sun - Ultra Sparc Enterprise 4000(trademark), SGI Challenge XL(trademark), and DEC 8400(trademark). Preferably, the server machine is configured for network 130 and supports TCP/IP protocol.

Depending upon the workstation employed, the operating system may be, for example, one of the following: Sun Sun OS 5.5 (Solaris 2 5), SGI IRIX 5 3 (or later), or DEC Digital UNIX 3 2D (or later).

Databases used in conjunction with this invention may be downloaded via a 4 X 4 Gb÷ FWSCSI 2, Fiber Link Raid Units 20Gb÷, or 4 DAT Tape Drive. A CD ROM drive may also be acceptable.

The client machine may be, for example, a Macintosh(trademark) (Apple Computer Inc. of Cupertino, CA), a PC, or a Unix workstation. It should also be TCP/IP capable with a Netscape or Internet Explorer Web Browser.

The network may include a 10Base-T, 100Base-T or higher connection, be TCP/IP capable, and provide access to Internet for HTML hyperlinks to external databases.

Figure 1C illustrates the accessibility of graphical viewer features in accordance with a preferred embodiment of the present invention. A graphical viewer in accordance with the present invention is preferably provided together with a suite of functions made available to users through a collection of user interface screens (e.g., HTML or Java(registered trademark) pages) viewed in the user interface of a biomolecular relational database. Typically, the interface will have a main viewer page from which various lines of query can be followed. In a preferred embodiment, the main viewer page (and other graphical viewers) are Java(registered trademark)-based applets running on the network system. Given the func

tionalities described herein, one of ordinary skill in the art would be able to implement the graphical viewers of the present invention in Java (registered trademark) or other programming environments. The viewer page is typically accessed from another page provided as part of the user interface of a biomolecular sequence relational database in connection with which the graphical viewer is used.

For example, a user interface screen (e.g., HTML page) 170 displays textual information relating to a plurality biomolecular sequences. One or more sequences displayed in the page 170 may be selected, for example, using the pointer provided in the GUI, to access another page 180 which displays additional information about the selected sequences. This page 180 may include a button which when selected accesses a main graphical viewer page 190. The graphical viewer page (e.g., Java(registered trademark) page) 190 graphically depicts information about the selected sequences. The page also preferably includes buttons 192 which allow a user to modify the graphical display. The buttons 192 may also include buttons which a user may select to access additional graphical viewer pages 194, 196, which graphically or otherwise display additional information relating to the graphically displayed sequence information in page 190.

Gene Locus Implementation

The invention will now be described with reference to a particular preferred implementation of the invention to graphically depict gene locus information. The invention will be described with reference to a database optimized for microbial data, such as that described in United States Patent No. 5,970,500. However, application of the present invention is by no means so limited. For example, the invention covers graphical viewers used in connection with databases optimized for other sources of biomolecular sequence data, such as animal sequences (e.g., human, primate, rodent, amphibian, insect, etc.) and plant sequences.

As noted above, a graphical viewer in accordance with the present invention is preferably provided together with a suite of functions made available to users through a collection of user interface screens viewed in the user interface of a biomolecular relational database. A main viewer page is typically accessed from another page provided as part of the user interface of a biomolecular sequence relational database in connection with which the graphical viewer is used, in this case a microbial genomic database. Figure 2 depicts one such other page from the microbial genomic database. The Contig Results page 200 displays a list of loci (identified by their LocusIDs) for genes localized to a particular "contig" (a group of assembled overlapping sequences), contig SAU1c0039, of the genomic sequence of a microbial organism, in this case Staphylococcus aureus.

By clicking on a particular LocusID in Contig Results page 200, a user accesses a Locus Information page, such as depicted in Figure 3. Clicking on the LocusID SAU100241 in page 200, returns the Locus Information page 300 which displays details about the locus SAU100241. The page also displays a Graphical Viewer button 302 which when selected launches a graphical viewer in accordance with the present invention.

Figure 4A depicts a main graphical viewer page 400 accessed by selecting the Graphical Viewer button 302 in Locus Information page 300. In this preferred embodiment, the graphical viewers are Java(registered trademark)-based applets that provide a graphical representation of a portion of a contig and its related loci. A graphical viewer in accordance with the present invention preferably includes a plurality of separate component viewers. Where more than one component viewer is featured it is preferably displayed in a single frame in order to enhance the effectiveness with which the graphically displayed data is conveyed to the user. A preferred embodiment includes three component viewers displayed in a sin

gle frame.

Thus, the graphical viewer 402 of page 400 has three viewer component panels on a single screen. The top panel 410 features a "legend viewer" 412, which shows the entire portion of the genome under consideration. The middle panel 430 features a "target viewer" 432, which allows a user to focus ("zoom in") on areas of the genome portion of particular interest. The bottom panel 450 features a "sequence depth viewer" 452, which contains graphical information illustrating the depth of coverage over the length of the genome portion represented in the target viewer 422.

The graphical viewer page 400 also includes several buttons and windows along the top 403 of the page 400 for accessing and displaying additional information. A menu bar 404 is also provided for accessing pull-down menus listing various command and control functions. A scale 415, 435, 455 depicted at the bottom of each viewer panel 410. The use of these features will be described in further detail below.

The legend viewer 412 always shows the entire portion of the contig which was loaded by the viewer when the user selected a contig in the previous screens. In a preferred embodiment, the viewer will load a predetermined default number of base pairs of the contig sequence. If the contig is shorter than the default", the entire contig will be depicted and the default will be adjusted . For example, in this embodiment, the viewer loads 30,000 base pairs starting at the first locus in the list on the Contig Results screen 200 (identified by its Hit ID), g2462967. The number of base pairs shown and the position on the contig may be determined with reference to the scale 415 depicted at the bottom of the legend viewer panel 410. The default value may, of course, be changed to any desired number.

The legend viewer 412 graphically represents contig SAUIc0039 as a line 414 which starts at coordinate (base pair number) 4467 and extends up to

coordinate 34,467, as may be seen with reference to the scale 415. The contig depicted in the viewer is identified in a ContigID window 407. In addition, the starting coordinate for the portion of the contig depicted by the legend viewer 412 (namely, the starting coordinate of the selected locus g2462967: 4467) is noted in the Start Coord window 405. These windows 405, 407 may also be used to enter information in order to control the information depicted by the viewer, as described further below. A user may bring upstream or downstream portions of the contig into view in the legend viewer 412, and the other component viewers, by clicking on the directional buttons 406.

In addition to the contig, the legend viewer 412 shows the various loci residing on the portion of the contig. The manner in which these loci are depicted illustrates the power of a graphical viewer in accordance with the present invention in presenting information in a highly effective manner.

The loci are represented by arrows 416. Each loci is located beside the contig line 414 according to its position on the contig and the direction in which it is read. The arrowhead represents the direction in which a locus is read. Loci which are read in the forward (+) direction are depicted above the contig line 414. Loci which are read in the reverse (-) direction are depicted below the contig line 414. In addition, other graphical features may be used to convey information about the graphically depicted loci. For example, loci for which the sequences obtained are above an established confidence threshold may be depicted as broken arrows.

In this preferred embodiment, the loci are also represented in different colors based on their protein's function. Proteins are grouped into various functional categories, with each category being assigned a color. For example, in this preferred embodiment, the proteins corresponding

tc loci are grouped according to the following categories/colors: Motil ity/Light blue; Virulence/Red; Transport/Light Green; Regulation/Magenta ; Macromolecule metabolism/Yellow; Small molecule metabolism/Dark blue; Structure/Dark Green; and Unclassified/Black. Of course other categorie s and colors may also be used. These arrow and color representation fea tures for loci are used in both the legend viewer and the target viewer, discussed below.

The target viewer 432 initially displays the same scope as the legend vi ewer 412. The scope of the target viewer may be modified, however, by c licking on the Zoom buttons 409. The Zoom In button provides a closer v iew of a portion of the contig shown in the legend viewer 412. The clos er view is depicted in the target viewer 432, with the scale 435 adjusti ng to reflect the amount of the zoom. The Zoom Out button provides a br oader view of the contig, up to the maximum of the default base pair num ber selected for the legend viewer (minimum magnification). The Zoom No ne button automatically returns to the minimum magnification.

Another way provided by a graphical viewer in accordance with the presen t invention to focus on a portion of interest of a contig 414 depicted i n the legend viewer 412 is to provide an outline, such as a colored (e.g ., red) box, around the portion of the contig 414 which is shown in the target viewer 432. In this preferred embodiment, when a red box surroun ds the entire legend viewer panel, the target viewer also display the en tire 30,000 base pairs. This is the situation illustrated in Figure 4A. When the Zoom buttons 409 are used, as described above, the red box is adjusted accordingly.

An area on the contig may also be zoomed into by direct user adjustment of the red box (known as "rubber banding"). The scope of the red box ma y be changed by clicking at a location in any of the viewer panels and d ragging the cursor with a mouse to another location. The red box will t

hen encompass the region between those two points, and only this region will be visible in the target and sequence depth viewers. Figure 4B depicts an updated page showing the viewer 402 after a user has zoomed in on the portion 434 of the contig 414 depicted in the legend viewer 412 between about the coordinates 14,200 and 18,200. The scale 435 at the bottom of the target viewer 432 has been adjusted to reflect the new scope of the zoomed target view.

Another feature of the target viewer is the loci are annotated. As may be seen in Figures 4A and 4B, annotations 436 are provided for loci arrows which are long enough to accommodate the annotation information. If a loci of interest is too short to be display its annotation, a user may zoom in further on the locus until it is long enough to allow the annotation to be displayed in the graphical representation.

Individual loci in the target viewer 432 may be selected for further analysis by clicking on the graphically depicted locus. A selected locus is highlighted in some manner, for example, by displaying a colored (e.g., red) box around its representation. Details about this locus my be viewed by double-clicking on the locus representation. Double-clicking opens a Selected Object Details window, such as depicted in Figure 5A. The Selected Object Details window 500 includes information about the locus, including its LocusID, gene (functional) category, base pair range, the sequence's homologous matches (preferably the number of homologous matches returned is limited to a preset number; for example, the top five matches are returned here) against other sequence databases, for example, the genpept database, and other information useful to researchers and relating to other features of the database system with which the graphical viewer is used. Many of the fields of information provided in the window 500 may be hyperlinks to other HTML pages or other screens.

The Selected Object Details window 500 includes an Alignment button 502.

Clicking on this buttor accesses an alignment viewer which provides a graphical representation of the locus sequence and its homologous matches. An example of an alignment viewer 510 in accordance with a preferred embodiment of the present invention is shown in Figure 5B. The alignment viewer 510 has three panels. The top two panels 512 and 514 provide a graphical representation of the locus identified in Figure 5A (SAU101156). The third panel 516 provides graphical representations of the five homologs noted in Figure 5A. The alignment viewer page also includes a number of buttons 518 which may be used to control the graphical representations. In particular, the page has Zoom buttons 520 which may be used to zoom into the sequence level of loci depicted in the lower two panels 514 and 516 (while the upper panel 512 maintains the depiction of the entire locus). Figure 5C illustrates this Zoom feature where the upper panel 512 has a colored box 522 around the portion of the locus depicted with its homologs at the sequence level in the two lower panels 514 and 516. In this embodiment, the amino acid sequences are shown. In other embodiments, the corresponding nucleotide sequences may also be shown.

An additional feature of the graphical viewer page 400 that becomes useful when the scope of the view in the target viewer 432 is focused in on a portion of the contig sequence shown in the legend viewer 412 is a scroll bar 470 at the bottom of the page. The scroll bar 470 allows a user to move along the portion 434 of the contig 414 to bring upstream or downstream portions of the contig 414 into view in the target viewer 432. The third panel 450 of the graphical viewer 402 in this embodiment of the present invention is the sequence depth viewer 452. The sequence depth viewer 452 provides a graph illustrating the depth of coverage, that is, the number of times that a given portion of the contig has been sequenced, over the length of the contig. The sequence depth viewer 452 disp

lays its graph for the contig or portion of the contig displayed in the target viewer 432. Thus, in Figure 4A, were the target viewer 432 and l egend viewer 412 have the same scope. the sequence depth viewer 452 disp lays a graph showing the depth of coverage over the 30,000 base pairs of the contig 414 from coordinates 4467 to 34,467, as indicated by the sca le 455 at the bottom of the sequence depth viewer panel 450. In Figure 4B, however, the sequence depth viewer 452 displays a graph showing the depth of coverage over the approximately 4000 base pairs of the portion 434 of the contig zoomed in on in the target viewer from about coordinat es 14,200 to 18,200, as indicated by the adjusted scale 455. The sequen ce depth viewer also includes a second scale 456 on the y-axis indicatin g the number of sequencing passes represented by the graph.

The manner in which this depth of coverage information is depicted provi des a further illustration of the power of a graphical viewer in accorda ance with the present invention in presenting information in a highly eff ective manner. A user of the graphical viewer is able to very quickly, at a glance, assimilate useful information relating to the confidence to be attributed to the sequence information depicted in the other panels of the viewer. In this preferred embodiment of the present invention, t he sequence depth viewer 452 depicts coverage as a sequence distribution graph 454. A particular advantage of this way of depicting of the dept h of coverage information is that it is particularly effective for clear ly providing this information in a graphical format which makes a clear visual impression and renders the data easily quantifiable, with referen ce to the y-axis scale 456. The coverage data for various regions is al so easily compared in this format.

In other embodiments of the invention, a sequence depth viewer may graph ically depict depth of coverage information in other ways. For example, the actual sequences from which the contig was assembled may be depicte

d. This way of depicting the sequence coverage information may provide useful information for some users who are concerned with the data acquisition process, for example, used in the contig's formation.

As noted above, the graphical viewer page 400 includes several buttons and windows along the top 403 of the page 400 for accessing and displaying additional information. Several of these have already been discussed, including the Start Coord 405 and ContigID 407 windows. Figures 6 and 7 illustrate additional features of a this embodiment of a graphical viewer in accordance with the present invention.

In addition to displaying the start coordinate for the contig sequence displayed in the legend viewer 612, the Start Coord window 605 may receive an entry from a user of a different starting coordinate. The entry of a different start coordinate will bring a different portion of a contig's sequence into view in the legend viewer. For example, Figure 6 shows a graphical viewer page 600 with the same settings as page 400, except that 0 has been entered in the Start Coord window 605. As a result, the contig sequence 602 and associated loci 604 shown in the legend viewer 612 is shifted 4467 base pairs upstream to the beginning of contig SAU1c 0039. The 4467-most downstream base pairs in the depiction of the contig 414 in Figure 4 are no longer visible in the viewer of page 600. The corresponding views are also depicted by the target viewer 632 and the sequence depth viewer 652.

Also, in addition to identifying the contig depicted in the viewer 402, the ContigID window 407 may receive an entry from a user of a different ContigID. The entry of a different ContigID will cause the default number of base pairs (preferably starting from the coordinate 0) of the contig sequence associated with the new ContigID to be loaded from the database associated with the viewer and displayed. For example, Figure 7 shows a graphical viewer page 700 with the ContigID SAU1c0016 entered in t

he ContigID window 707. As a result, the contig sequence 702 and associated loci 704 shown in the legend viewer 712 are that for contig SAU1c0016. The corresponding views are also depicted by the target viewer 732 and the sequence depth viewer 752.

As also noted above, the graphical viewer 400 includes a menu bar 404 for accessing pull-down menus listing various command and control functions. The File pull-down menu lists standard commands found in applications software packages such as save and print, etc. The Edit pull-down menu provides a list of categories for editing the parameters of the graphical viewers, including the default contig sequence length display number and the colors used to represent various features in the viewers.

Of particular interest is the View pull-down menu which, together with allowing the user to select which features should be included in the various viewer displays, also includes a View Sequences on Coords 804 option. A graphical viewer page 800 is shown in Figure 8A with the View pull-down menu 802 selected. Selection of the View Sequences on Coords 804 option from the menu 802 accesses a page 810 listing the sequences used to assemble the contig depicted in the graphical viewer 402, together with the coordinates of each sequence which indicate its coverage. Selecting a sequence from the list, such as the second one in the list, 806503 054F1 (5201,5690) 812, and clicking the Sequence Database button 814 accesses a database of the raw sequences used to assemble contigs in the database system associated with the graphical viewer and returns a Sequence Retrieval Results page 820, depicted in Figure 8C. The Sequence Retrieval Results page 820 depicts the actual nucleotide sequence 822 of the sequence 812 selected in Figure 8B.

A generalized process by which a graphical viewer system in accordance with a preferred embodiment of the present invention returns graphical representations of gene locus information to a user is depicted in Figure

9. This process flow shows only some of the main features of a preferred embodiment of the present invention in order to illustrate in process flow form some of the options for graphically displaying sequence data in accordance with an embodiment of the present invention. It is not intended to provide a comprehensive depiction of the present invention.

The process 900 begins at 901 and at a step 902 data for a selected locus and its associated contig are loaded into the graphical viewer. As noted above, the locus may be selected from a list in a HTML page provided as part of the user interface of a biomolecular sequence relational database in connection with which the graphical viewer is used, in this case a microbial genomic database. At a step 904, a graphical display of the selected locus on its contig is provided. Preferably, the graphical display has a plurality of components for representing different aspects of the biomolecular sequence data associated with the selected locus. In a particularly preferred embodiment depicted in Figures 4A and 4B and described above, the graphical representation is a viewer having three components: a legend viewer, a target viewer, and a sequence depth viewer.

If no further entries or zoom adjustments are made, the process may end at step 940 following the graphical display at step 904. However, a user may want to use the graphical viewer to extract and display additional information relating to the selected locus or other loci, and the viewer provides additional functionalities for this purpose.

The graphical representation of the data displayed by the graphical viewer may be modified in a variety of ways. Also, additional information may be accessed by selecting various objects (namely, loci) in a viewer.

For example, a user may enter a new contigID in a field provided in a graphical viewer page, such as window 407 in Figure 4A. If so, decision step 906 is answered in the affirmative and the new contig and its loci

are graphically depicted in a viewer at a step 914. A user may also enter a new start coordinate, such as in Start Coord window 405 in Figure 4A. If so, decision step 908 is answered in the affirmative and the graphical display is adjusted to show the contig in the new coordinate range at a step 916. In addition, as described above, a user may choose to focus in on a particular portion of a graphically depicted contig. If so, decision step 910 is answered in the affirmative and the graphical display in the target viewer, in this embodiment, is adjusted to show the contig in the zoomed view at a step 918. If any of these decision steps are answered in the negative the graphical viewer display remains unchanged.

After any of these actions, or in the alternative, a user may select an object to obtain further information. In a preferred embodiment, the loci depicted in a target viewer component of the graphical viewer may be selected by clicking on its representation. If so, decision step 920 is answered in the affirmative and the depiction of the locus in the target viewer may be highlighted with a colored box. If a user wishes to obtain detailed information about the selected loci, the user may do so by double clicking on the depiction of that locus. If so, decision step 921 is answered in the affirmative and a Java(registered trademark) page showing detailed information about the selected locus is shown at a step 922.

Another feature of this aspect of a preferred embodiment of the present invention is a graphical alignment viewer, as described above. A user may elect to display a graphical viewer which shows the alignment of the amino acid sequence of the loci of interest against some homologous sequences. If so, decision step 924 is answered in the affirmative and the alignment is graphically displayed in a graphical viewer at a step 926. A user may also be provided with the option of displaying a multiple org

anism viewer to view graphical representations of homologous and paralog ous loci of the locus of interest. For example, if decision step 920 is answered in the affirmative, a multiple organism viewer may be accessed at a step 930 when a decision step 928 is answered in the affirmative.

Further details of the operation of a multiple organism viewer in acco rdance with a preferred embodiment of the present invention are describe d below with reference to Figures 10A-10E and 11.

Of course the selected object details and multiple organism selection de cisions are independent of each other and could just as easily have been presented in other ways in Figure 9. Further, it should be noted that t he system allows the user to exit from the graphical viewer mode at any time. This option is not depicted in Figure 9.

A further option available for accessing further information from a grap hical viewer in accordance with the present invention is the display of actual nucleotide or amino acid sequences for a selected sequence associ ated with the locus of interest and its contig. In a preferred embodime nt, a user may choose this option by clicking on a button in a graphical viewer page such as depicted in Figure 4A. If so, decision step 912 is answered in the affirmative and a list of sequences sequence identifier s) and coordinates for the sequences from which the contig displayed in the viewer was assembled is displayed at a step 932. A user may then se lect a sequence from the list. If so, decision step 934 is answered in the affirmative and the actual nucleotide sequence (in this case) of the selected sequence is displayed. The process ends at 940.

As with other data displayed in graphical viewers in accordance with the present invention. the data used in this aspect of the invention is obt ained from an associated biomolecular sequence database and system. The organization and operation of such systems may vary. Examples are prov ided in the Incyte Pharmaceuticals patents previously referred to herei

n. Given the description of the functionality and displays herein, one of skill in the art would be able to implement the graphical viewer of the present invention in any such system.

Multiple Organism Viewer

As noted above, a graphical viewer in accordance with the present invention may also provide for the graphical representation and comparison of multiple portions of the genome of one or more organisms based on a locus of interest and its corresponding paralogs (related loci from other portions of an organism's genome) and homologs (related loci from another organism's genome). A preferred embodiment of such a multiple organism viewer is described with reference to Figures 10A-10D, below.

Figure 10A depicts a main graphical viewer page 1000, like that shown in Figures 4A and 4B. In Figure 10A, a box ("rubber band") 1002 has been placed around a region 1004 of the portion of the contig 1006 displayed by the legend viewer 1008 component of the graphical viewer 1010. This region 1004 of the contig 1006 is displayed by the target viewer 1020 component of the graphical viewer 1010, and its coverage is depicted by the sequence depth viewer 1030 component. In the target viewer 1020, a box 1022 around locus SAU100242 indicates that that locus has been selected. As noted previously, the main viewer page 1000 includes a Multiple Organisms button 1001.

Clicking on the Multiple Organisms button 1001 when a locus has been selected in the target viewer retrieves from the database associated with the viewer and displays a list of all libraries containing homologs and paralogs of the selected locus. Figure 10B depicts a window 1040 showing a list of libraries retrieved for the locus SAU100242 selected in the previous page shown in Figure 10A. To access a list of individual homologs and paralogs, a user may select one or more libraries in the list displayed in this window 1040. Clicking on the Multiple Organisms button 1

042 retrieves the individual homologs and/or paralogs and displays them.

Figure 10C depicts an example of a window 1045 showing a list of homol ogs and paralogs for the locus SAU100242 from the libraries selected in screen 1040 shown in Figure 10B. A hit description for each locus is al so provided.

A user may then choose to produce a graphical display of the originally selected locus (e.g. SAU100242) and the selected homologous and paralogo us loci displayed in the list of Figure 10C. By clicking on the Multipl e Organisms button 1046 in windowl045, the locus of interest and its hom ologs and paralogs are loaded into a multiple organism viewer in accorda nce with a preferred embodiment of the present invention, and the locus of interest and the selected homologs and paralogs are displayed. Figur e 10D depicts an example of such a multiple organism viewer page 1050. The multiple organism viewer page 1050 provides a single panel multiple organism viewer 1052 graphically depicting the selected locus of interes t (SAU100242) on its contig (SAU1c0039) and the selected homologous and paralogous loci on their respective contigs. Figure 10D, shows a viewer 1052 graphically displaying five (5) contigs in a single page: SAU1c00 39 1061, PRT1c0129 1062, SAU2c0391 1063, SEP1c0220 1064, and SHA1c0122 1 065. Contig 1061 is shown together with its loci, including the selecte d locus SAU100242 1071 depicted in bold and italicized in order to more clearly identify it. In the embodiment depicted in Figure 10D, the loci are annotated with a hit description rather than a Locus ID. Each of t he other contigs is also depicted with its loci alongside, and with the loci homologous to SAU100242 (respectively, loci 1072, 1073, 1074 and 10 75) shown in bold italics.

The multiple organism viewer 1052 illustrates another example of the pow er of a graphical viewer in accordance with the present invention to con vey biomolecular sequence information in an effective way. As noted abo

ve, the selected locus and its homologous and paralogous loci may be sho wn bold and italicized, or in other type, such as a particular color, in order to distinguish them as the loci of interest. As may also be seen in Figure 10D, the loci of interest for the graphically displayed conti gs are aligned in the page 1052 so that a visual comparison of adjacent loci on the various contigs is easily achieved. This visual representat ion may be further enhanced through use of the complement feature descri bed below with reference to Figure 10E.

Further features of such a graphical viewer in accordance with this em bodiment of the invention may be accessed by clicking on pull-down menu selections 1053 provided in the multiple organism viewer page 1050. The menu selections include File, View and Help selections that provide fea tures such as described above with reference to Figure 4A. The Show sel ection 1054 accesses a list of all of the loci listed in the window illu strated in Figure 10C and loaded into the multiple organism viewer. By selecting a locus from the Show pull-down menu, a user may determine tha t the locus along with the contig on which it resides will be displayed or hidden. clicking on the loci, a user may determine that a locus will be displayed or hidden. The Show menu may also provide for the same de termination to be made with respect to the contigs.

The Complement menu selection 1055 allows a user to manipulate the gra phical representations of the contigs and loci in order to facilitate th e extraction of salient information from the data. In particular, the c omplement menu selection 1055 allows the user to perform a reverse compl ement on any of the contigs displayed in the multiple organism viewer 10 52. In this way, the homologous and paralogous loci displayed in the vi ewer 1052 may be shown in the same reading direction so that a user may more easily locate patterns of related loci adjacent to the loci of inte rest. Figure 10E depicts a multiple organism viewer page 1080 in which

the loci of interest depicted in page 1050 are shown with the same reading direction by use of the complement feature to show the reverse complement of contigs 1062, 1063 and 1065.

Shortcuts for the complemeot feature, as well as other features described herein, may be made available to a user according to methods well known to those of skill in the art. For example, the complement of a locus (contig) may be shown by holding down the shift key on a keyboard used to interface with the computer system on which the graphical viewer in operating while clicking oo the contig.

Figure 11 depicts a flow chart for a generalized process of the operation a multiple organism viewer in accordance with a preferred embodiment of the present invention. The process 1100 starts at 1101, and at a step 1102 the multiple organism viewer system receives a selection of a locus of interest, for example by clicking on a locus in the target viewer of Figure 10A. At a step 1104, a list of libraries containing loci homologous or paralogous to the selected locus of interest is displayed in a window. This display may be initiated by a user clicking on a button, such as the Multiple Organisms button in Figure 10A. Next, the system receives a selection of one or more libraries from the list at a step 1106, and at a step 1108 a list of loci from the selected libraries which are homologous or paralogous to the selected locus of interest is displayed in a window. At a step 1110, the system receives a selection of loci homologous or paralogous to the selected locus of interest from the list to be displayed. Then, at a step 1112, the selected loci and their respective contigs are graphically displayed in a multiple organism graphical viewer. In a preferred embodiment, the viewer shows all of the contigs and loci in a single panel in order to facilitate comparison of the graphically depicted data. The process ends at 1114.

Implementation

It is important to note that the present invention may be implemented as a system or a method, and may be embodied on a variety of computer-readable media that include program instructions, etc. for performing various operations described herein. As noted above, the system implementation is preferably in association with a biomolecular sequence relational database system. The method is a computer-implemented method, generally involving the operation of such a system. The media may be any computer-readable media. Examples of computer-readable media include, but are not limited to, magnetic media such as hard disks, floppy disks, and magnetic tape; optical media such as CD-ROM disks; magneto-optical media such as floptical disks; and hardware devices that are specially configured to store and perform program instructions, such as read-only memory devices (ROM) and random access memory (RAM). The invention may also be embodied in a carrier wave travelling over an appropriate medium such as airwaves, optical lines, electric lines, etc.

Conclusion

Although the foregoing invention has been described in some detail for purposes of clarity of understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims. It should be noted that there are many alternative ways of implementing methods, media and systems of the present invention. As noted previously, the scope of the invention is not limited to use with a microbial genomic database system such as that in connection with which the invention is primarily described above. Given the description provided herein, one of skill in the art would understand how to use the present invention in connection with a variety of computer-based biomolecular sequence database systems. For example, a graphical viewer in accordance with the present invention may be used in connection with database systems employed to store and analyze other types and forms of nucleic acid

sequences or expressed nucleic acid or amino acid sequences. Accordingly, the present embodiments are to be considered as illustrative and not restrictive, and the invention is not to be limited to the details given herein, but may be modified within the scope and equivalents of the appended claims.

4. BRIEF DESCRIPTION OF DRAWINGS

Figure 1A is a block diagram of a client-server Intranet for providing database services in accordance with one embodiment of the present invention.

Figure 1B is a schematic representation of the various software documents and entities employed by the Figure 1A client-server Intranet to provide biological information in response to user queries.

Figure 1C is a block diagram illustrating the accessibility of graphical viewer features in accordance with a preferred embodiment of the present invention in connection with a biomolecular sequence database.

Figure 2 is a screen shot (HTML page) depicting a Contig Results page for a graphical user interface of a genomic sequences database suitable for selecting a locus to be viewed with a biomolecular sequence graphical viewer in accordance with one embodiment of the present invention.

Figure 3 is a screen shot depicting a Locus Information page for a graphical user interface of a genomic sequences database suitable for accessing a biomolecular sequence graphical viewer in accordance with one embodiment of the present invention.

Figure 4A is a screen shot depicting a main page of a biomolecular sequence graphical viewer in accordance with one embodiment of the present invention.

Figure 4B is a screen shot depicting a main page of a biomolecular sequence graphical viewer modified to illustrate the zoom feature in accordan

ce with one embodiment of the present invention.

Figure 5A is a Selected Object Details window in accordance with one embodiment of the present invention.

Figures 5B and 5C are screen shots depicting an alignment viewer in accordance with one embodiment of the present invention.

Figure 6 is a screen shot depicting a main page of a biomolecular sequence graphical viewer modified to illustrate the new starting coordinate feature in accordance with one embodiment of the present invention.

Figure 7 is a screen shot depicting a main page of a biomolecular sequence graphical viewer modified to illustrate the new ContigID feature in accordance with one embodiment of the present invention.

Figures 8A-8C are screen shots depicting pages of a biomolecular sequence graphical viewer illustrating a feature which displays an actual biomolecular sequence in accordance with one embodiment of the present invention.

Figure 9 is a flow chart depicting a process flow by which gene locus information may be viewed with a biomolecular sequence graphical viewer in accordance with a preferred embodiment of the present invention.

Figures 10A-10E are screen shots depicting the operation of a multiple organism biomolecular sequence graphical viewer in accordance with one embodiment of the present invention.

Figure 11 is a flow chart depicting a process flow by which multiple organism gene locus information may be viewed with a biomolecular sequence graphical viewer in accordance with a preferred embodiment of the present invention.

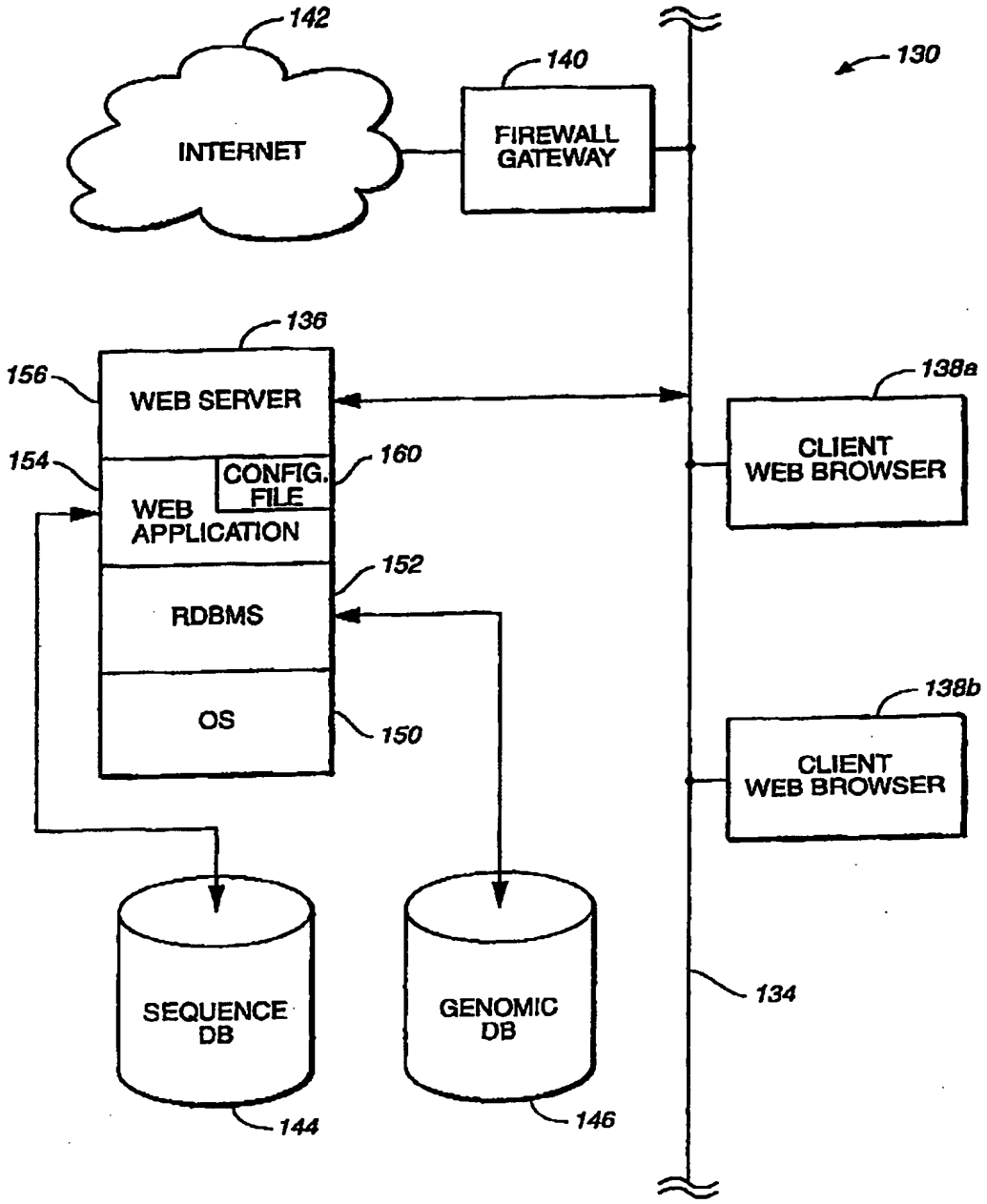

142
INTERNET
140
FIREWALL GATEWAY
130
136
156
WEB SERVER
138a
CLIENT WEB BROWSER
154
CONFIG. FILE
160
WEB APPLICATION
152
RDBMS
OS
150
138b
CLIENT WEB BROWSER
134
SEQUENCE DB
GENOMIC DB
144
146

FIG._1A

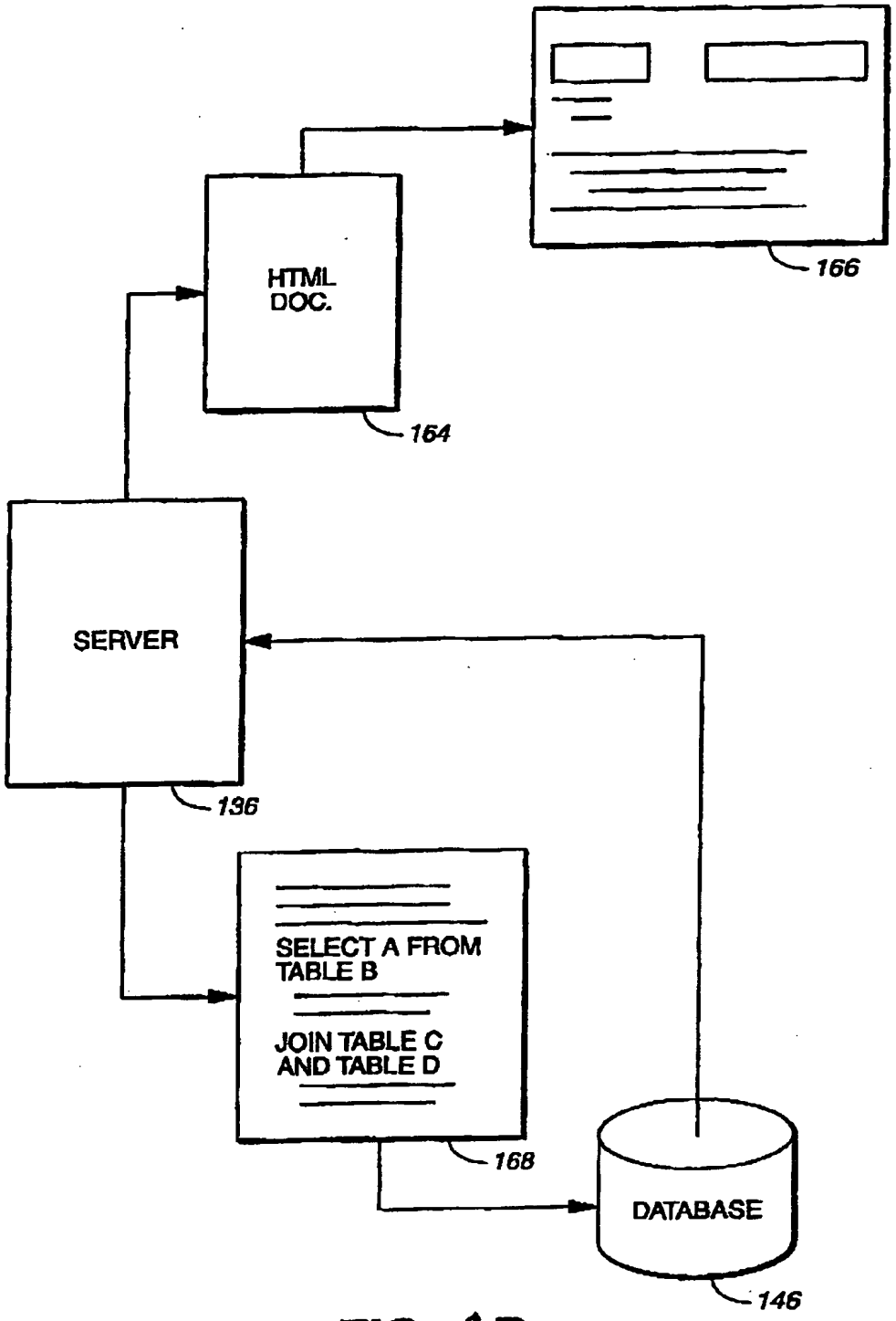

HTML
DOC.
164
166
SERVER
136
SELECT A FROM
TABLE B
JOIN TABLE C
AND TABLE D
168
DATABASE
146

FIG._1B

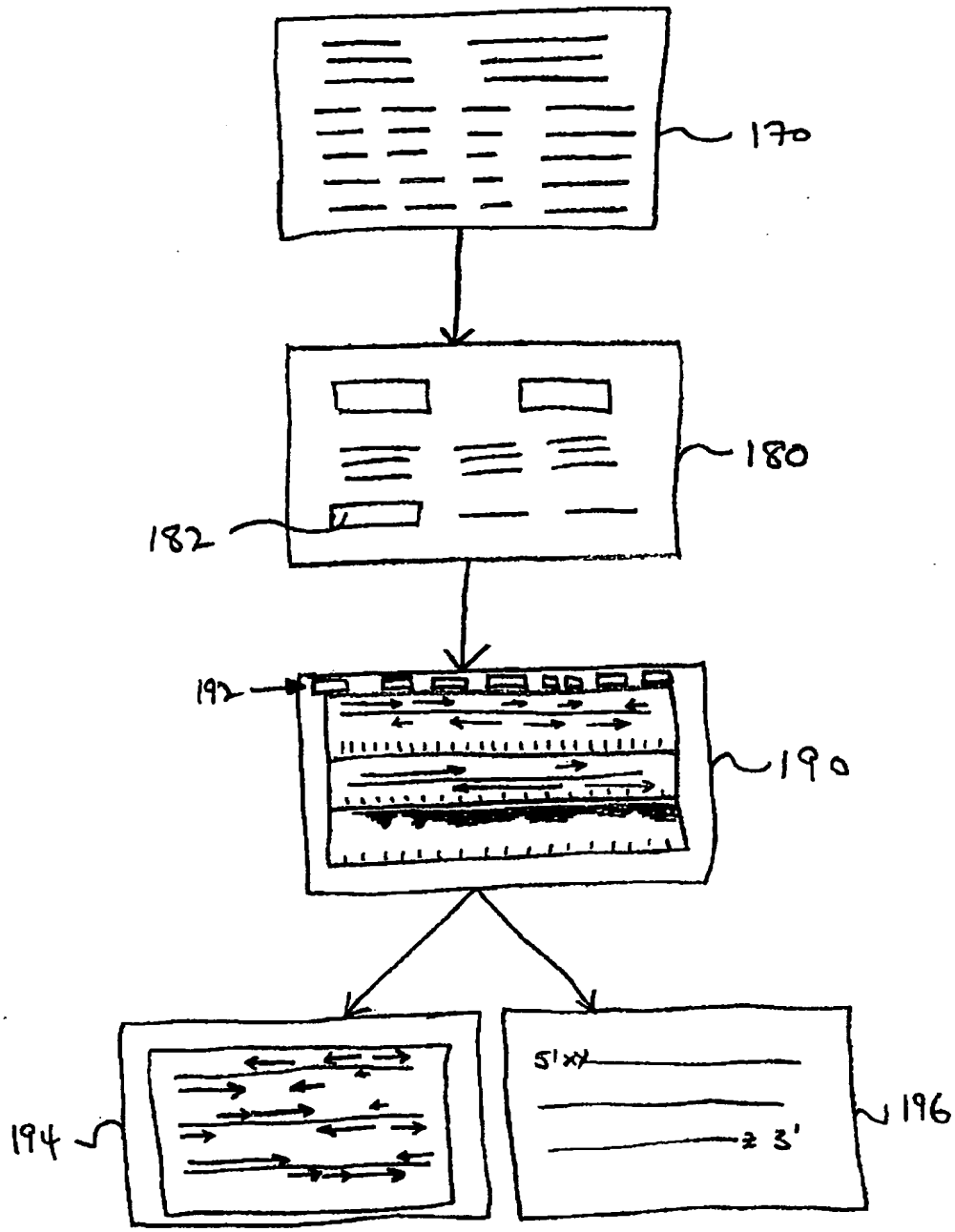
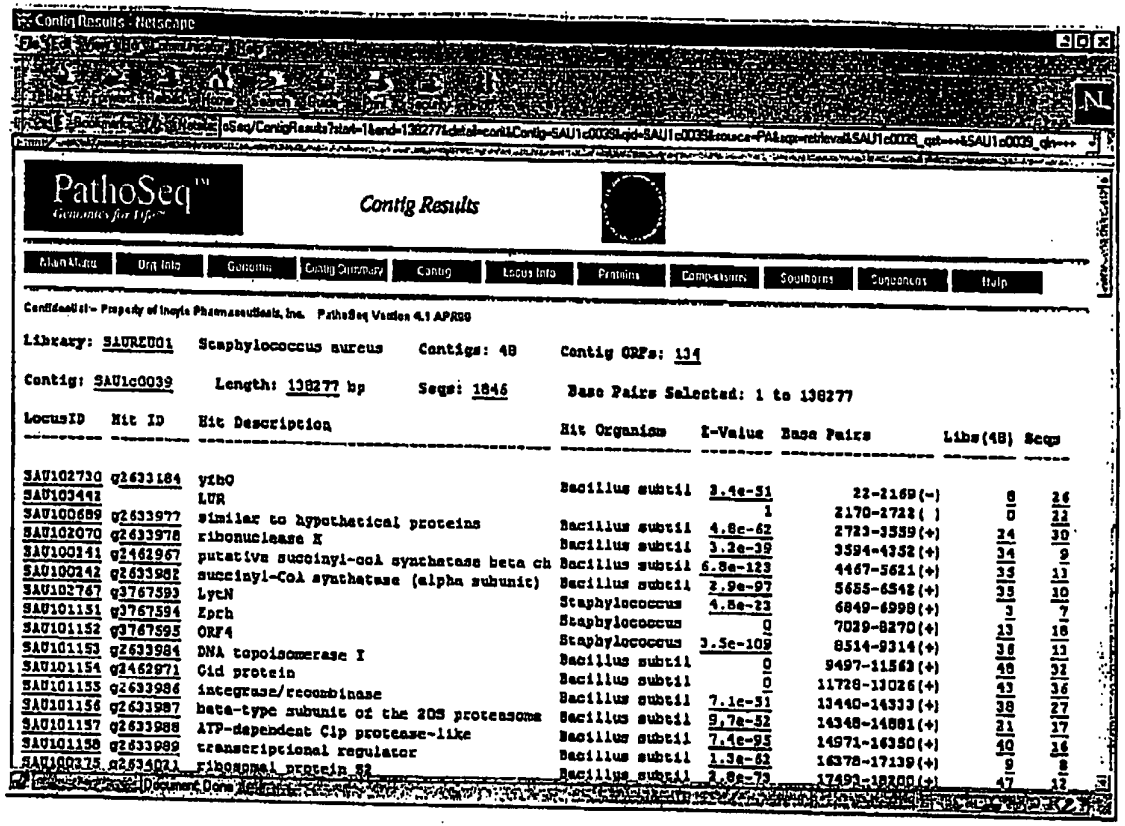


FIG. 1C



200

FIG. 2

特開2001-125929

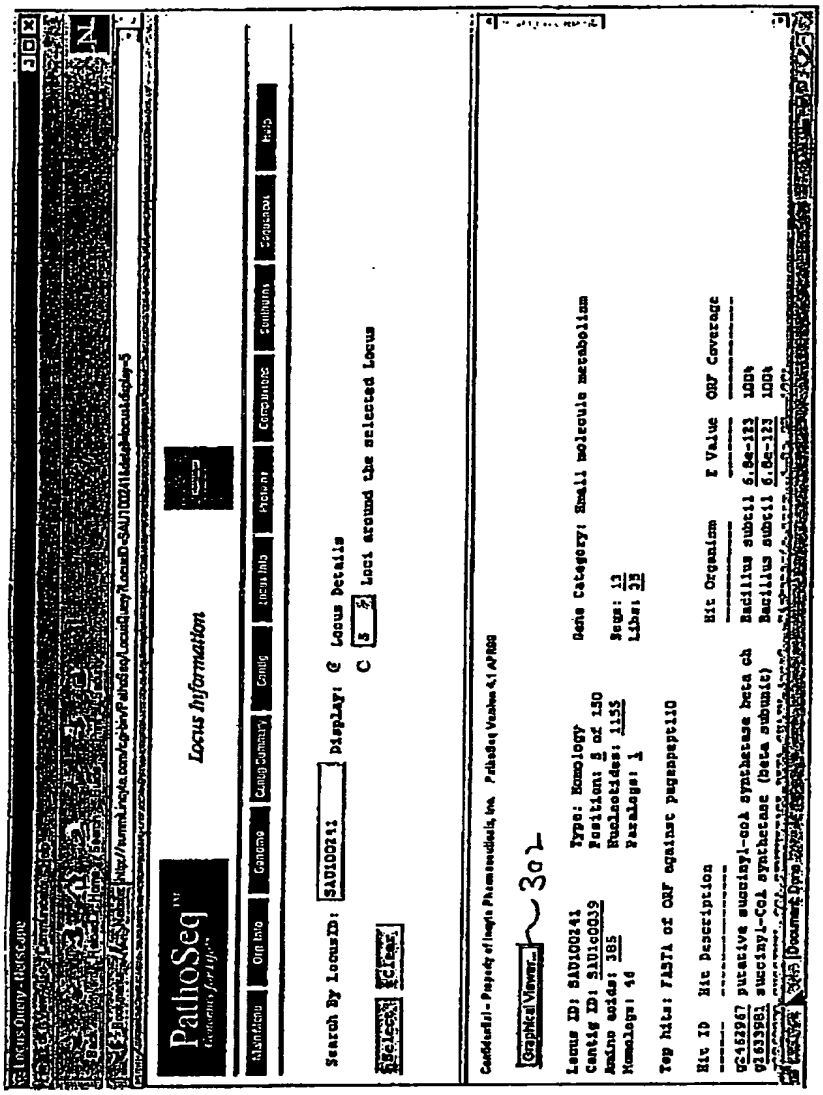

PathoSeq
Locus Information
Search By LocusID:
Display:
Locus Details
Loci around the selected Locus
Graphical Viewer
Locus ID: SAD1000341
Contig ID: SAUic0039
Amino acids: 385
Homologs: 46
Type: Homology
Position: 5 of 150
Nucleotides: 1155
Paralogs: 1
Gene Category: Small molecule metabolism
Segs: 13
Libs: 39
Top hits: FASTA of ORF against pepseptall0
Hit ID
Hit Description
Hit Organism
E Value
ORF Coverage
g2182967 putative succinyl-coA synthetase beta ch
Bacillus subtil
6.8e-123
100%
g1633981 succinyl-CoA synthetase (beta subunit)
Bacillus subtil
6.8e-123
100%
301
300

FIG. 3

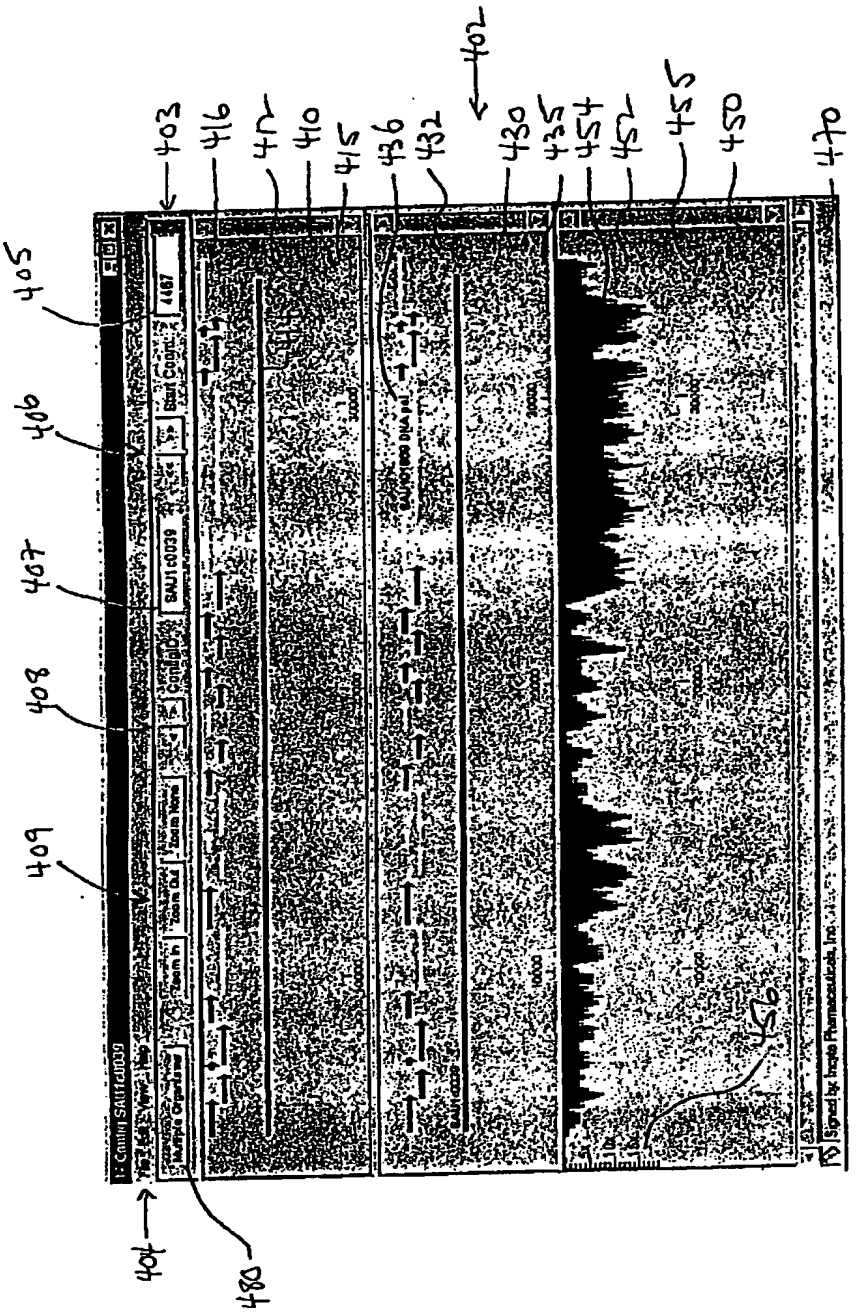

402
403
404
405
406
407
408
409
410
415
416
417
430
432
435
436
450
452
454
455
460
470
480

FIG. 4A

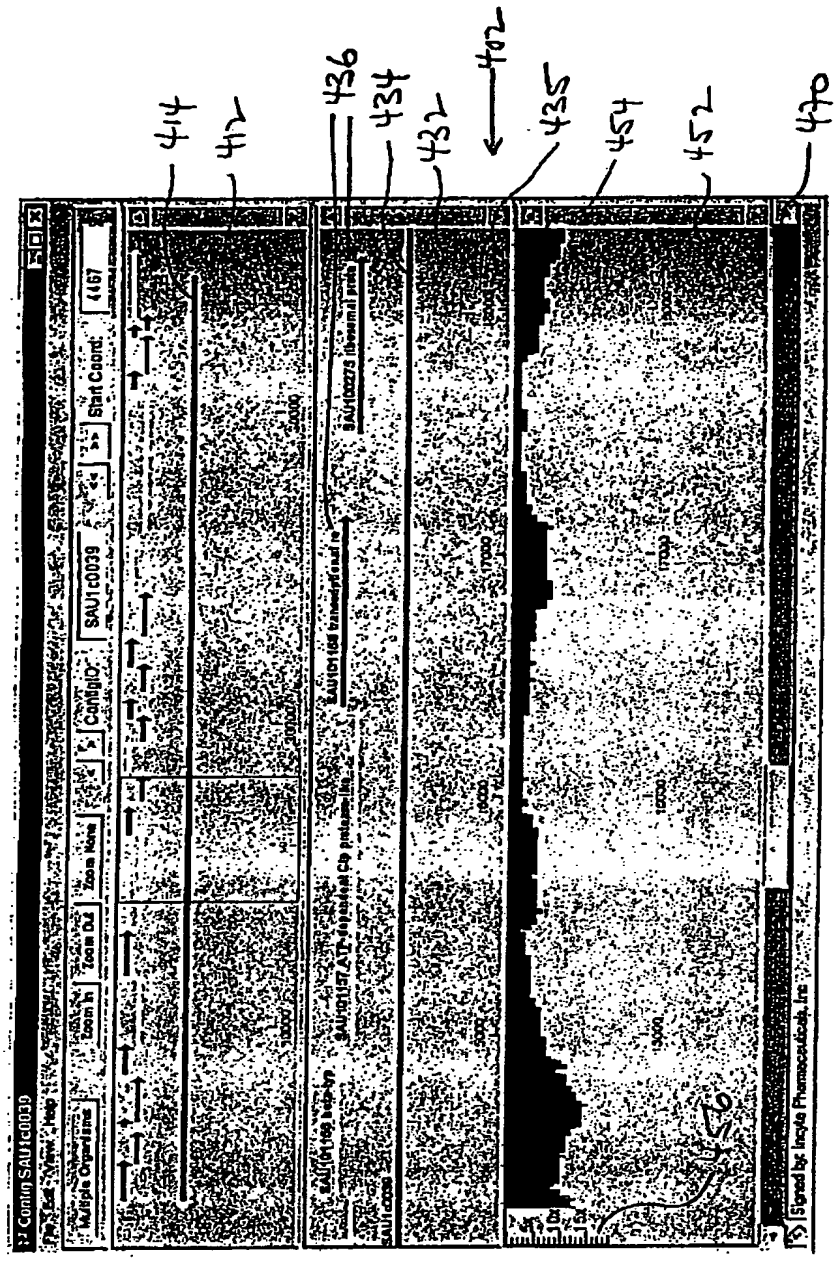


FIG. 48

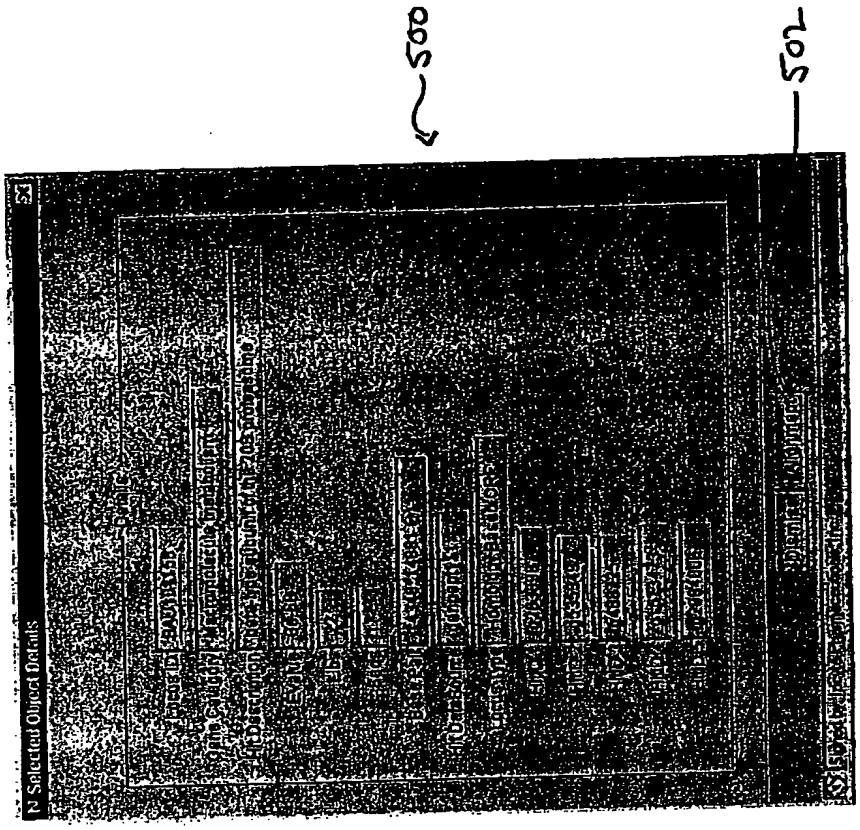


FIG. 5A

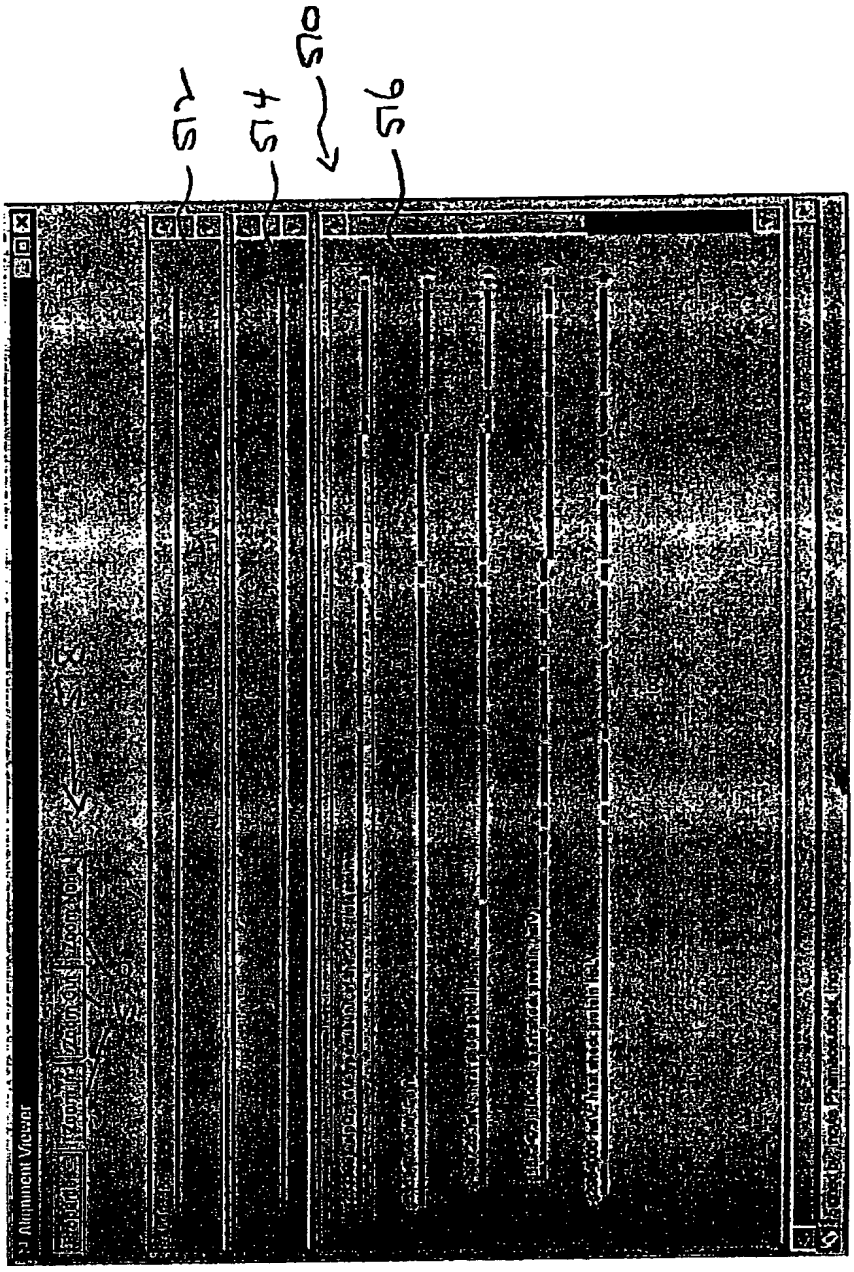


FIG. 5B

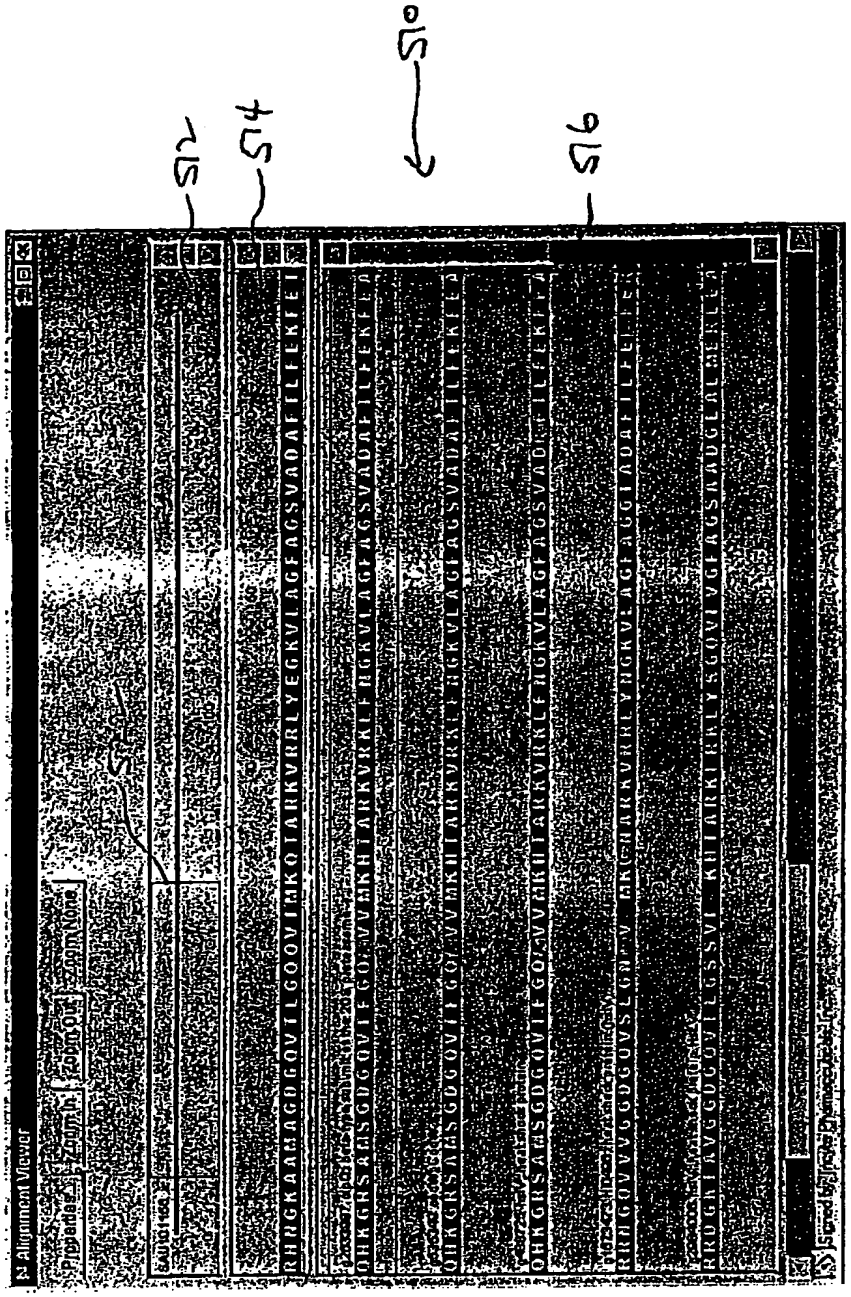


FIG. 5C

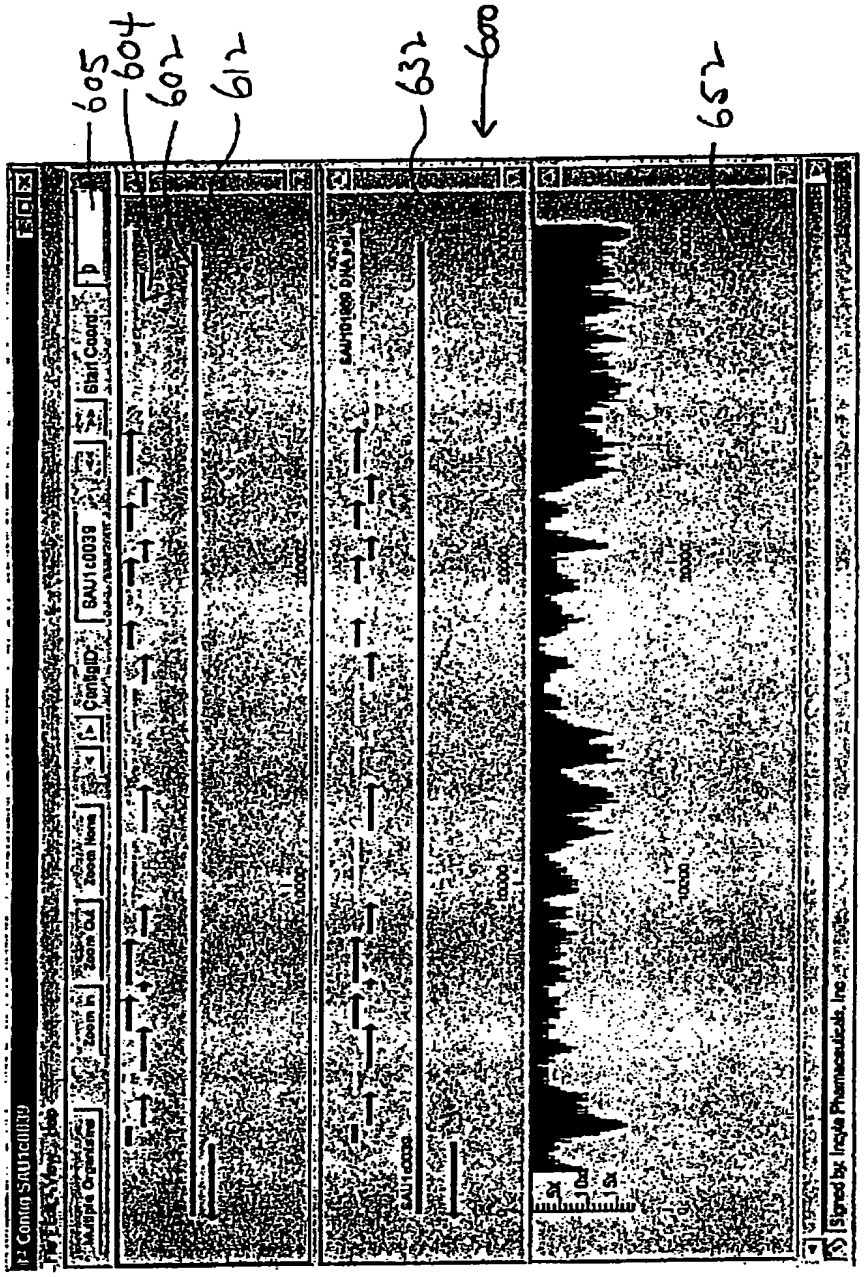


FIG. 6

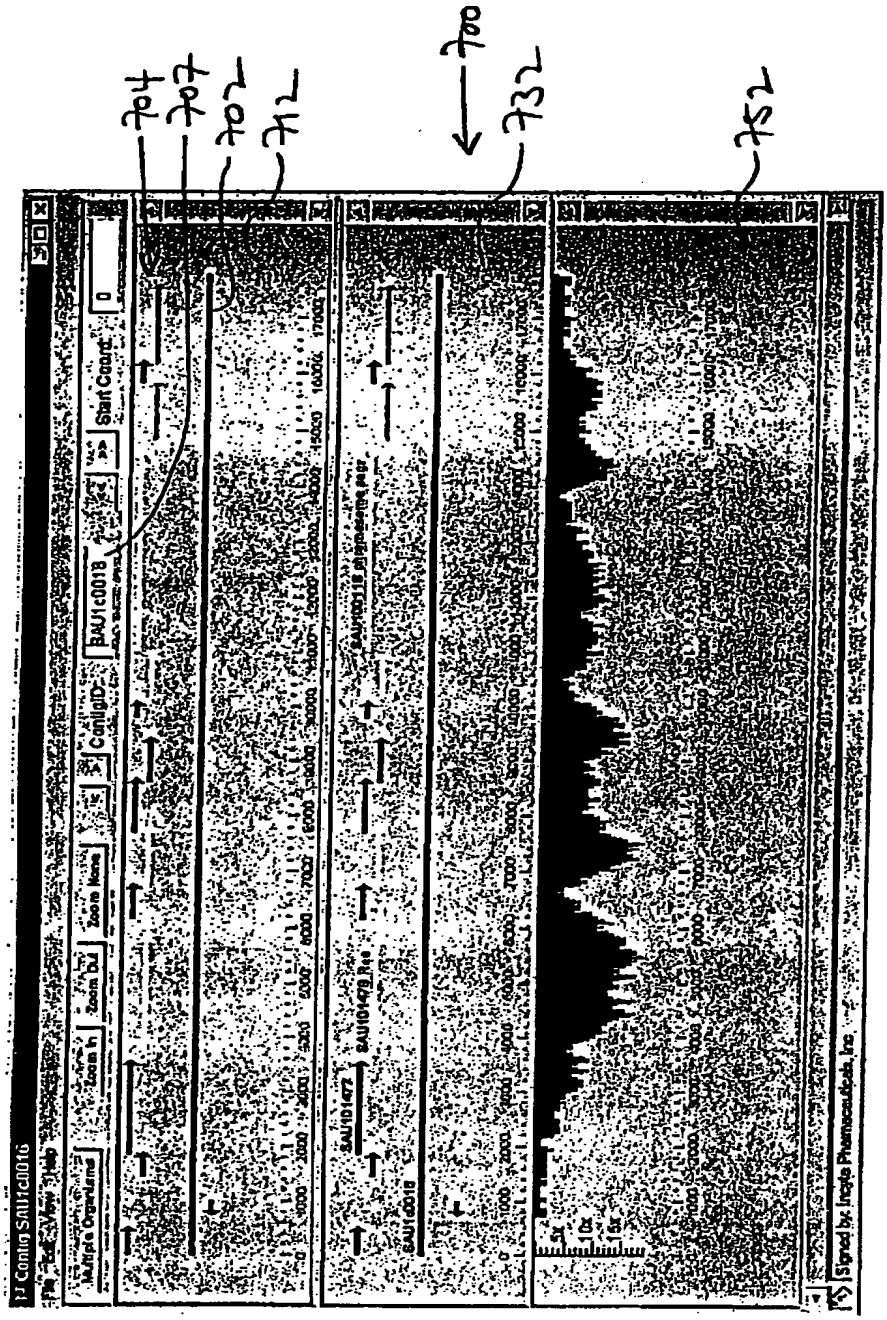

700
704
707
702
712
732
752
JJ Config SAU1c0016
File Edit View Help
Multiple Organisms
Zoom In
Zoom Out
Zoom None
Config ID:
BAU1c0018
Start Coord:
1
SAU1c0018
SAU101472
SAU101473
SAU1c0018
Signed by Incyte Pharmaceuticals, Inc.

FIG. 7

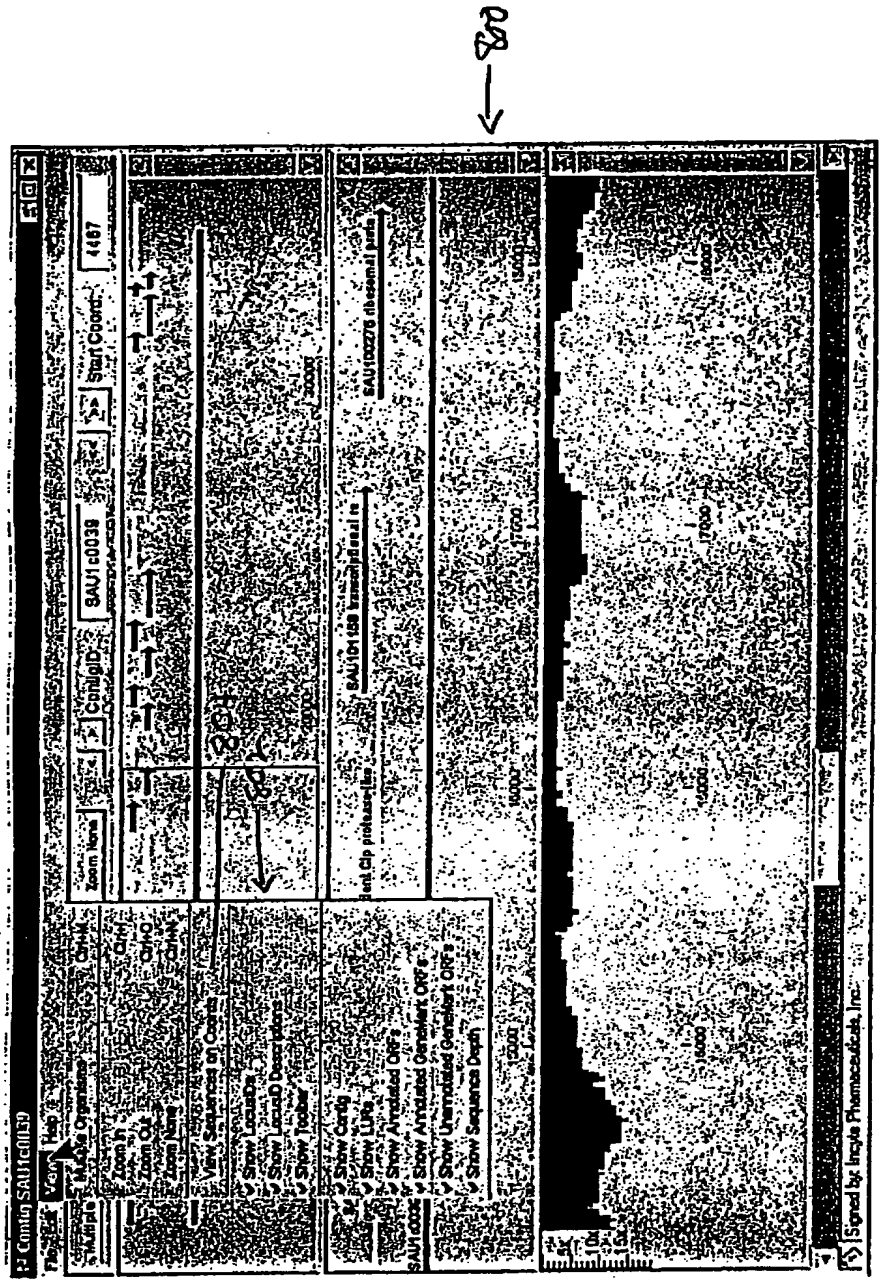

800
801
802
Config SAU1c0039
View
Help
Mode Organizer
Zoom In
Zoom Out
Zoom None
Show Locators
Show Locator Descriptions
Show Footer
Show Config
Show LUTs
Show Annotated ORFs
Show Annotated Gene/Mark ORFs
Show Unannotated Gene/Mark ORFs
Show Sequence Depth
ConfigID
SAU1c0039
Start Coord:
1487
Zoom None
Zoom In
Zoom Out
Signed by Incyte Pharmaceuticals, Inc.

FIG. 8A

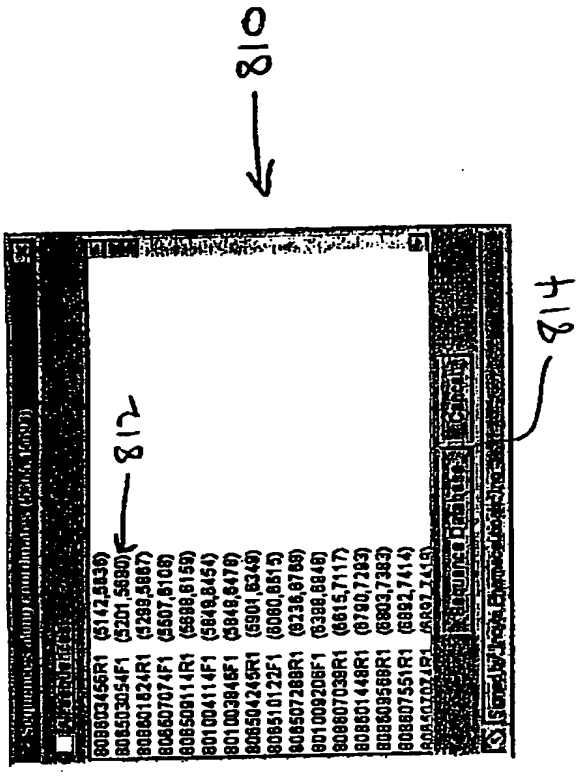

810
812
814
Sequences display coordinates (GSAS 1990)
Reading Display
Cancel
80803456R1 (5142,5836)
80803054F1 (3201,5880)
80801824R1 (4289,5867)
80607074F1 (3507,6108)
80809114R1 (5898,6159)
80104114F1 (5849,8454)
80100384F1 (5848,8478)
80804245R1 (5901,6340)
80851012F1 (6000,8815)
80807288R1 (6238,6788)
80100920F1 (6398,8846)
80807038R1 (6615,7117)
80801448R1 (6790,7283)
80809598R1 (6803,7383)
80807551R1 (6892,7414)

FIG. 88

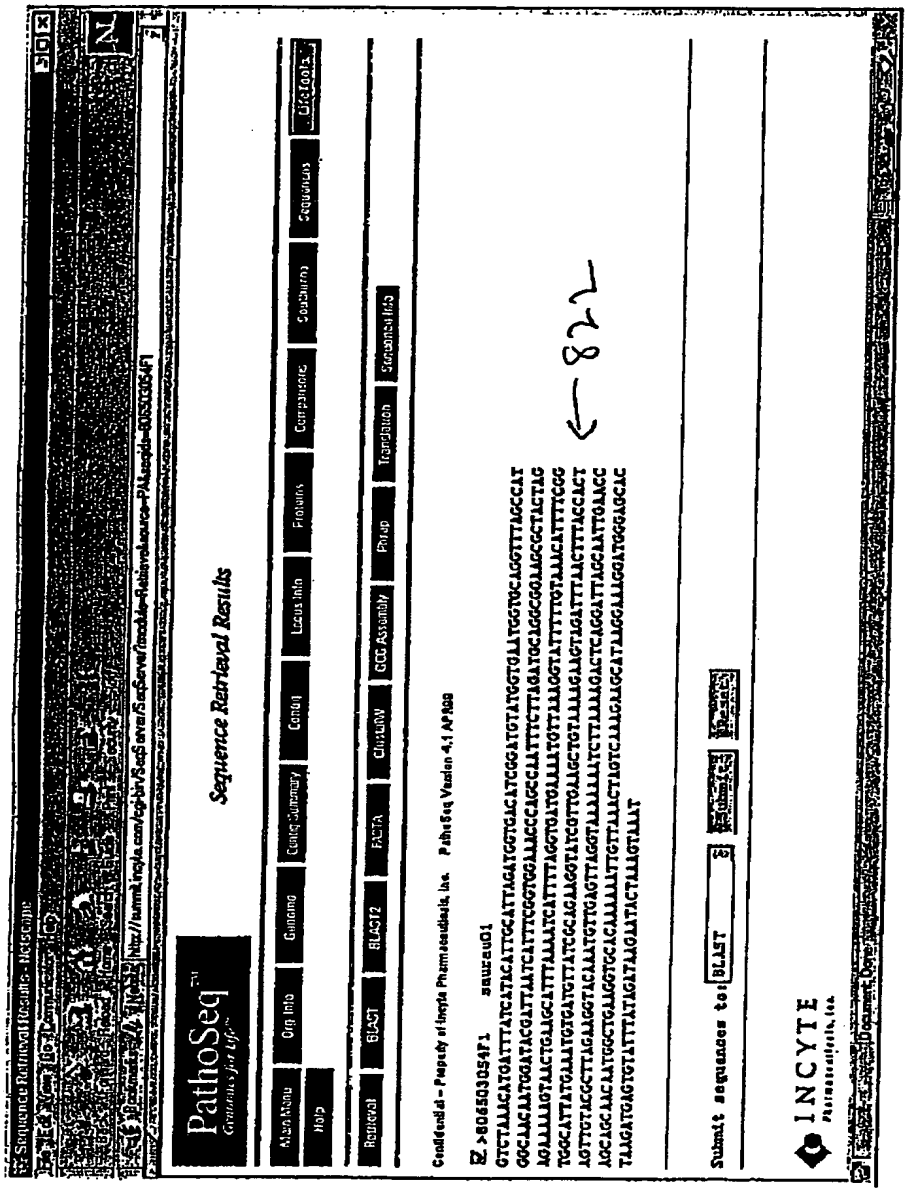

PathoSeq
Sequence Retrieval Results
Main Menu
Org Info
Contains
Comp Summary
ContigU
Locus Info
FindVariants
Comparisons
SeqSummary
Sequences
OrgTools
Help
Retrieval
BLAST
BLAST2
FASTA
ClustalW
CCC Assembly
PFAM
Translation
Structure Info
Confidential – Property of Incyte Pharmaceuticals, Inc. PathoSeq Version 4.1 AP500
>80650305471 sau2au01
Submit sequences to: BLAST
INCYTE
820
826

FIG. 8C

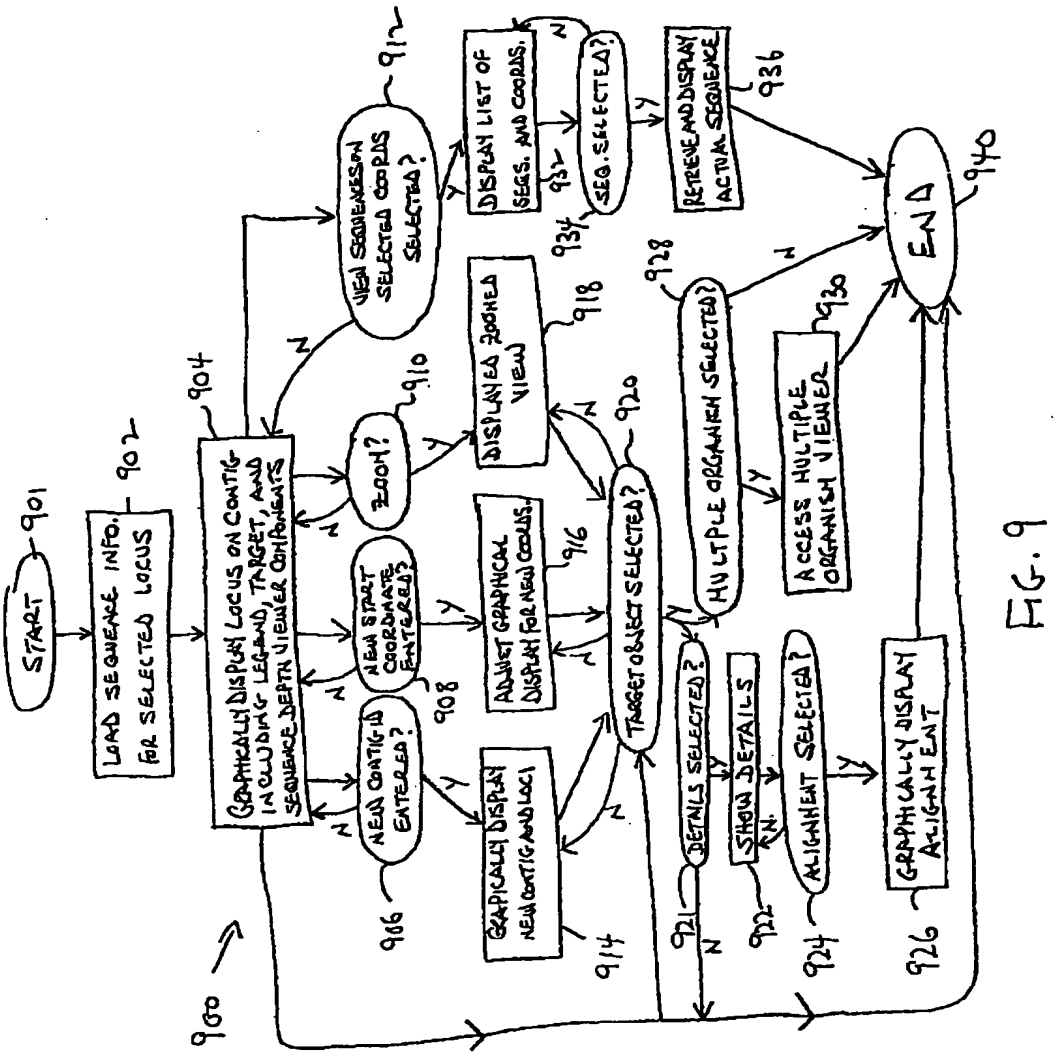

START
901
LOAD SEQUENCE INFO. FOR SELECTED LOCUS
902
GRAPHICALLY DISPLAY LOCUS ON CONTIG INCLUDING LEGEND, TARGET, AND SEQUENCE DEPTH VIEWER COMPONENTS
904
ZOOM?
910
DISPLAYED ZOOMED VIEW
918
NEW START COORDINATE ENTERED?
908
ADJUST GRAPHICAL DISPLAY FOR NEW COORDS.
916
NEW CONTIG ID ENTERED?
906
GRAPICALLY DISPLAY NEW CONTIG AND LOCI
914
TARGET OBJECT SELECTED?
920
VIEW SEQUENCES ON SELECTED COORDS SELECTED?
912
DISPLAY LIST OF SEGS. AND COORDS.
932
SEQ. SELECTED?
934
RETRIEVE AND DISPLAY ACTUAL SEQUENCE
936
MULTIPLE ORGANISM SELECTED?
928
ACCESS MULTIPLE ORGANISM VIEWER
930
DETAILS SELECTED?
921
SHOW DETAILS
922
ALIGNMENT SELECTED?
924
GRAPHICALLY DISPLAY ALIGNMENT
926
END
940
900

FIG. 9

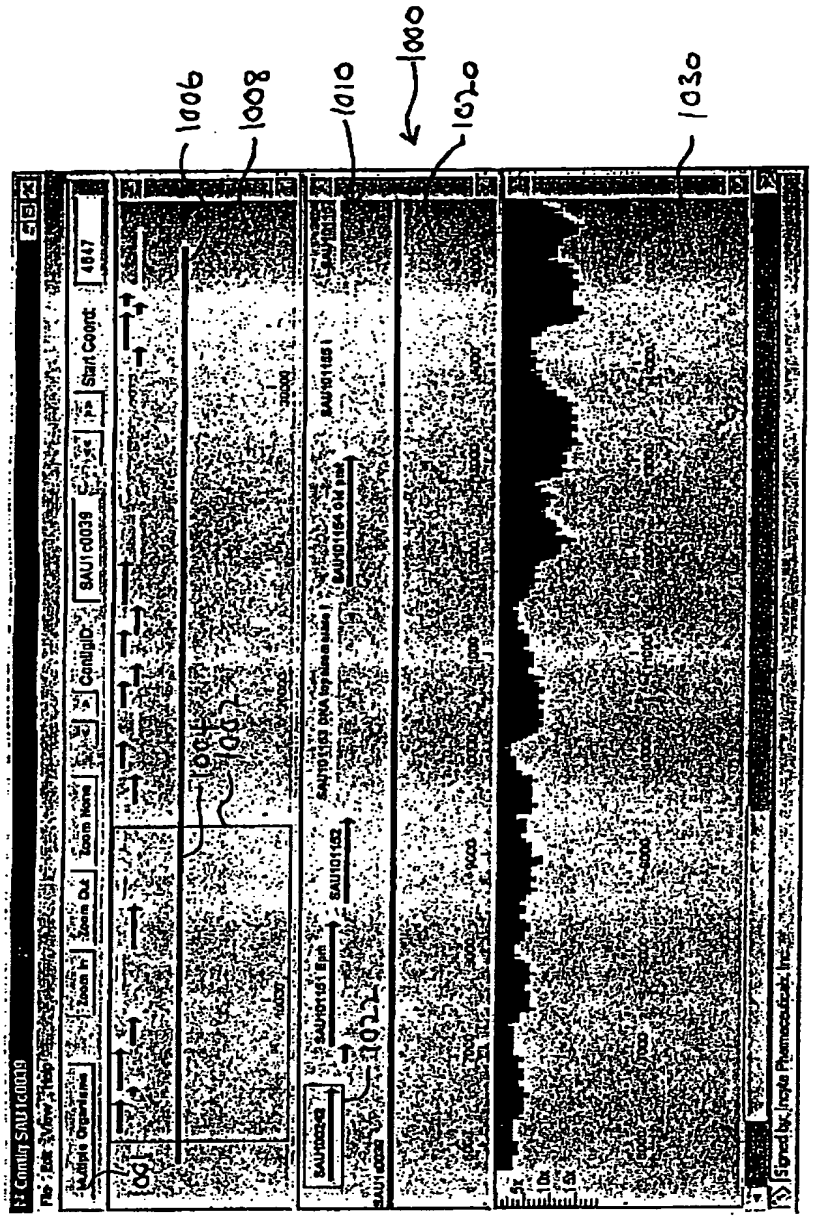

Contig SAU1c0039
File Edit View Help
Multiple Organisms
Zoom In
Zoom Out
Zoom Home
ContigID:
SAU1c0039
Start Coord:
4847
Signed by: Incyte Pharmaceuticals, Inc.
1000
1002
1004
1006
1008
1010
1020
1022
1030

FIG. 10A

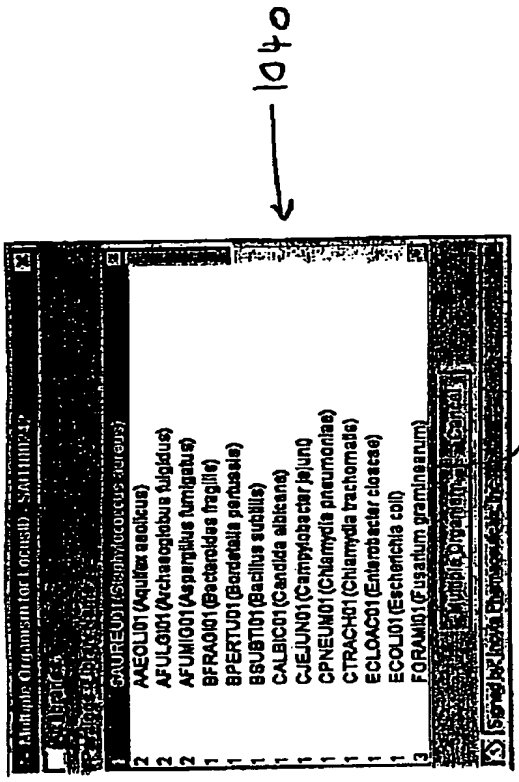


FIG. 10B

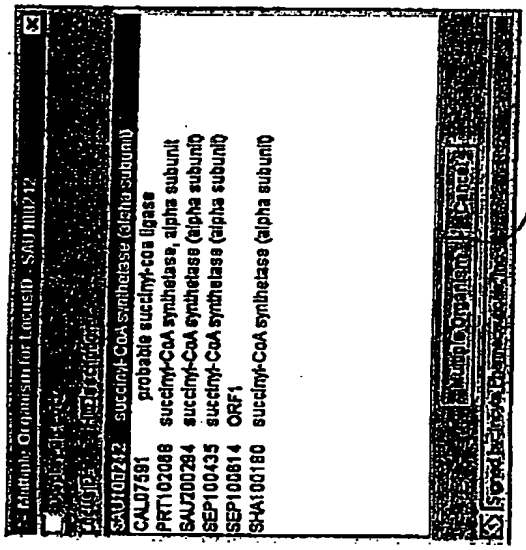


FIG. 10C

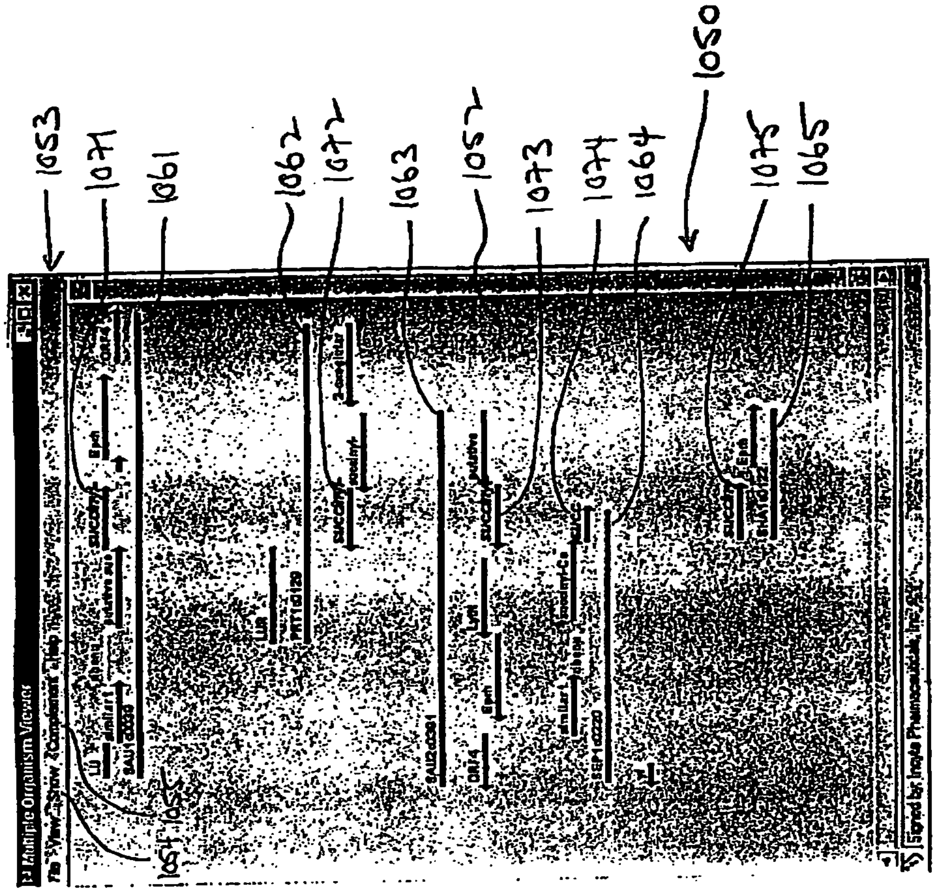

1053
1071
1061
1062
1072
1063
1052
1073
1074
1064
1050
1075
1065
1051

FIG. 10D

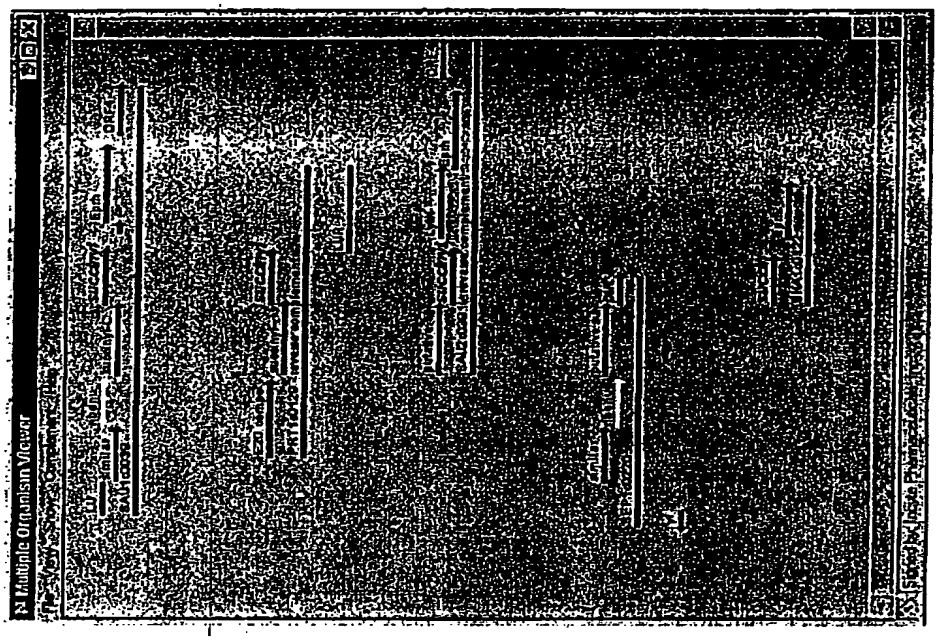


FIG. 10E

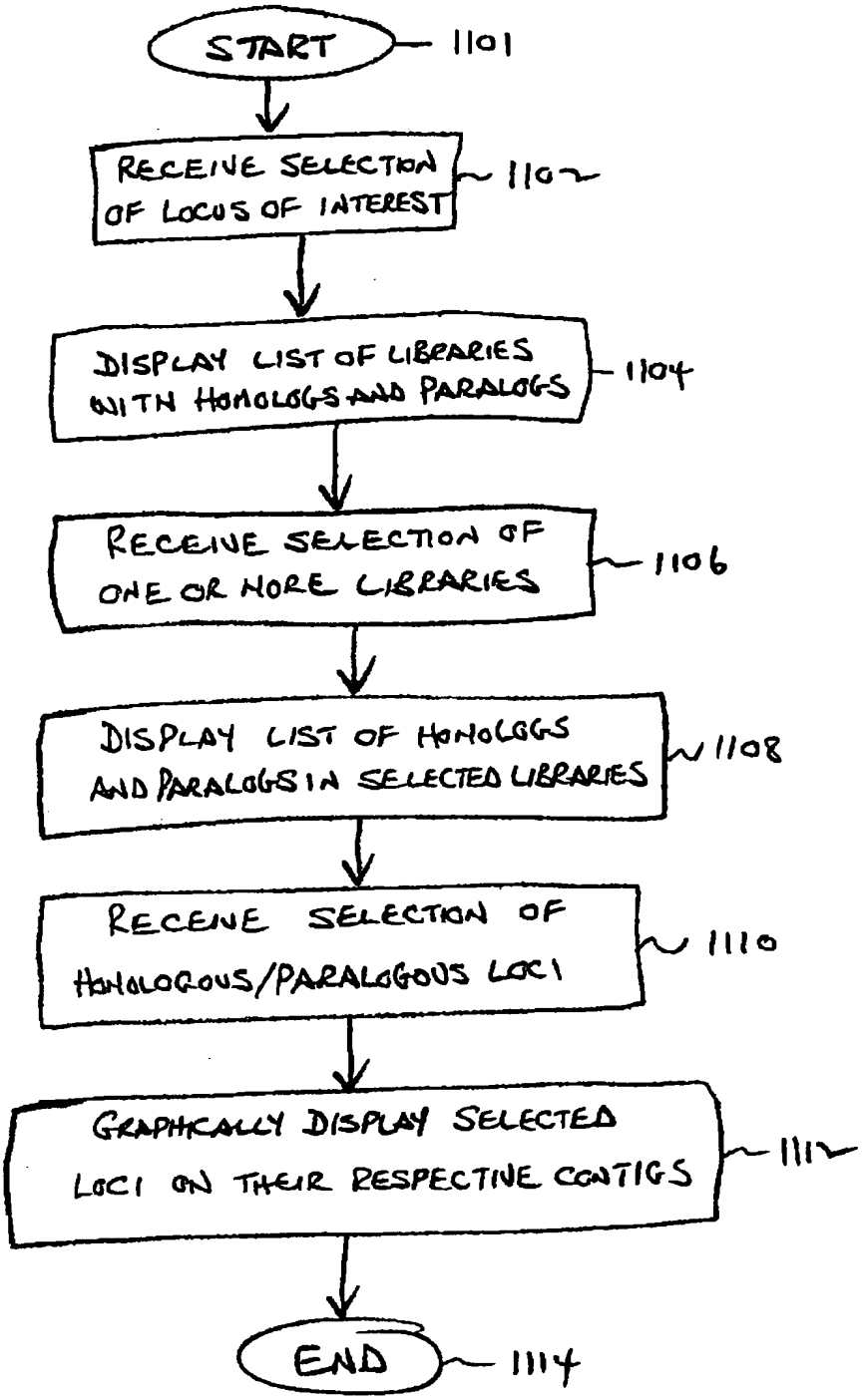

START
1101
RECEIVE SELECTION OF LOCUS OF INTEREST
1102
DISPLAY LIST OF LIBRARIES WITH HOMOLOGS AND PARALOGS
1104
RECEIVE SELECTION OF ONE OR MORE LIBRARIES
1106
DISPLAY LIST OF HOMOLOGS AND PARALOGS IN SELECTED LIBRARIES
1108
RECEIVE SELECTION OF HOMOLOGOUS/PARALOGOUS LOCI
1110
GRAPHICALLY DISPLAY SELECTED LOCI ON THEIR RESPECTIVE CONTIGS
1112
END
1114

FIG. 11

1. Abstract

Disclosed are methods, media and systems for graphically displaying comp uter-based biomolecular sequence information. Generally, biomolecular s equence information may be graphically depicted in a variety of differen t forms in accordance with the present invention. The sequence informat ion may be composed of nucleotide or amino acid sequence information or both. The graphical depictions may be in several different formats prov iding different information relating to the sequences, and may be displa yed in one or more screens of a computer user interface.

2. Representative Drawing

Figure 4A